THE UNIVERSITY
of EDINBURGH

Understanding *In Vivo* Modelling of Depression

By

Alexandra Bannach-Brown

THE UNIVERSITY
of EDINBURGH

Doctor of Philosophy

The University of Edinburgh

2018

Understanding *In Vivo* Modelling of Depression

By

Alexandra Bannach-Brown

AARHUS UNIVERSITET

Doctor of Philosophy

Aarhus University

2018

Abstract

Major Depressive Disorder (MDD) is the leading source of disability globally. Treatment-resistance among patients is common and even effective pharmacological therapies have a delayed effect on symptom relief. Better understanding of the mechanisms underlying depression and the search for potential effective and novel therapeutic targets are high research and healthcare priorities. Animal models are commonly used to mimic aspects of the phenotype of the human disorder to characterise candidate antidepressant agents. Despite these tools, no new pharmacological interventions have been discovered in the last decade and no reliable biomarkers have been identified for clinical use.

Systematically reviewing the literature on animal models of depression may provide an overview of our current understanding of the underlying biological mechanisms and why no new therapies have been effectively translated to clinic. This field of research is large, and over 70,000 potentially relevant articles were identified in 2016. Therefore systematically reviewing this literature presents challenges for human resources. To combat these challenges, the following contributions to the field have been made: (1) the novel application of machine learning techniques to identify errors in human systematic review citation screening; and (2), the novel application of regular expression dictionaries to large corpuses of preclinical animal literature to help cluster publications into the disease model investigated and drug intervention tested. These tools have been applied for systematic review and meta-analysis methodology to the field of animal models of depression.

All literature on animal models of depression has been systematically identified using searches carried out in PubMed and EMBASE in May 2016. This literature has been screened with the help of machine learning classification algorithms, based on a random set of dual human screened records (5749 records). This achieved a sensitivity of 98.7% and a specificity of 86% as assessed on in an independent validation dataset.

Machine learning has been used to identify human screening errors in the set of documents used to train the algorithm. Correction of these errors with further human intervention, sees an improvement in specificity to 88.3%. These algorithms allow irrelevant documents to be automatically removed, reducing the corpus to 18,407 articles that highly likely to be relevant to the research area of animal models of depression. Custom-made regular expression dictionaries of (1) techniques to induce depressive-like phenotypes in animals, and (2) known antidepressants have been curated. The text-mining dictionaries for anti-depressant drugs and commonly used methods of model induction have been applied to categorise and visualise this large corpus of records to allow prioritisation of sub-topics of depression for further in depth systematic review and meta-analyses. These machine-assisted tools for systematic review methodology are available free to use, online.

Systematic review and meta-analysis has been conducted on two sub-topics of the literature on animal models of depression. Firstly, the literature on the effects of ketamine as an anti-depressant in animal models of depression has been summarised with systematic review techniques and the effects of ketamine on depressive-like behaviour in the forced swim test, has been pooled using meta-analysis. The timing of administration of ketamine relative to the outcome assessment was significantly associated with decreases in effect size. This meta-analysis revealed no statistically significant heterogeneity between the studies. Secondly, the literature on use of gut microbial altering interventions to induce and treat depressive-like phenotypes in animal models of depression has been summarised and their effects have been pooled across studies using meta-analysis. The systematic review and meta-analysis of microbiota interventions identified a broad range of outcomes investigated in the primary literature and several probiotic treatments to reduce depressive-like behaviour were investigate gaps in the literature. Finally, a primary hypothesis-confirming animal experiment, where measures to reduce the risk of bias have been implemented was carried out to investigate the effects of prebiotics on depressive- and anxiety-like behaviour in a genetic animal model of depression, the Flinders Sensitive Line (FSL) rats.

Online tools have been developed to provide an overview of animal models of depression and anti-depressant drugs investigated in the literature, using systematic

review methodology and automation tools. This thesis reports meta-analyses on two sub-topics within animal models of depression; the effect of microbiota interventions, and the effects of ketamine; along with a primary animal experiment to test the effects of prebiotics on depressive-like behaviour in a genetic rodent model of depression.

Lay Abstract

Major Depressive Disorder (MDD) is a psychiatric condition characterised by low mood and lack of interest in hobbies and activities once enjoyed. Depression is the leading source of disability globally. Not all patients respond to current drug treatments and these drugs have a delayed effect on symptom relief. It is therefore a healthcare priority to find effective treatments and to better understand the underlying biology behind the symptoms and the development of depression. A considerable number of studies performed in animals are used to investigate potential novel drugs before the treatments are licensed for use in humans. Many different interventions are used to 'mimic' symptoms of human conditions in animals, commonly referred to as animal models.

Systematic review is a methodology to systematically identify and summarise all literature in a particular topic. Meta-analysis of the data can be subsequently performed across many studies to see under which circumstances a drug is effective. This study aims to provide an overview of the literature on animal models of depression.

Searching online medical databases picked up many studies that likely contain information about using animals to model depression. To see whether all studies identified (70,365 studies) are relevant requires a great deal of human resources. Here, tools to help automate this process have been tested and implemented successfully. An online tool has been developed to provide an overview of all literature reporting animal models of depression, categorised by different types of antidepressant drugs investigated and by different techniques used to induce the model of depression.

Using these tools, two systematic reviews of sub-topics within this field have been carried out. Firstly, a systematic review and meta-analysis of all literature on the use of ketamine, was conducted, providing a quantitative summary of the effect of this drug across all the studies. Further, a systematic review of interventions targeting the gut bacteria was carried out to provide an overview. The results of the primary studies in this review were also summarised quantitatively with meta-analysis. Results from

this review revealed that there were still many questions left to be answered by this literature about different types of interventions on the gut bacteria.

A primary animal study was conducted to expand our understanding of the effect of different microbiome-targeting interventions on depression. A genetic animal model of depression was used. Animals received non-digestible fibres that promote the growth various gut bacteria. The effect of this was measured in behavioural tests of depression and anxiety.

Online tools have been developed to provide an overview of animal models of depression and different anti-depressant drugs investigated, using systematic review methods and tools to automate this process. This thesis reports quantitative summaries of two sub-topics within animal models of depression; the effect of gut microbiota-targeting interventions, and the effects of the anti-depressant ketamine. This thesis reports a primary animal experiment to test the effects of prebiotics on depressive-like behaviour in a genetic rodent model of depression.

The tools of systematic review and meta-analysis provide an overview of the evidence available on potential treatments that have been tested in animals; how different exposures such as stress and genetic alterations affect the development of depression; the underlying biological mechanisms behind these treatments and exposures; and the quality of this research can help us make decisions about future research. We can use this research to better design future animal experiments, and together with other research, make decisions about drugs to investigate with clinical trials in humans. In future, further tools can be developed to help automatically synthesise and summarise evidence from many studies so that this process can bring knowledge

Dansk Resume

Depression er en af de mest alvorlig psykiske sygdomme i verden og er forbundet med massive samfundsøkonomiske omkostninger. Antidepressiv medicin kan være længe om at virke og omkring 50% af patienterne er behandlingsresistente. Der er derfor behov for øget viden omkring biologiske markører, som muligvis har indflydelse på sygdommens udvikling. Dyremodeller af depression kan bruges til at identificere biologiske reaktionsveje, hvilket kunne bidrage til en mere målrettet diagnosticering. Anvendelse af dyremodeller har også til formål at kunne give indikationer om præparater, som kan udvikles til forbedret og individualiseret anvendelse af antidepressiv medicin.

Formålet med mit ph.d. projekt var at belyse publicerede forskningsresultater for at skabe et bedre overblik og større indsigt i forskningen i dyremodeller af depression. En systematisk review er en metode, som systematisk indsamler og analyserer videnskabelig litteratur vedrørende et specifikt emne. Dermed forbedres mulighederne for at identificere de nøgleområder, herunder biomarkører og mulige antidepressiv stoffer, som kan danne fundamentet for bedre diagnosticering og behandling af depression.

Vores systematiske analyse har identificeret 70.365 unikke videnskabelige studier relateret til dyremodeller af depression. Det har sine udfordringer med så stort et datamateriale. Idet en manuel gennemgang af datamateriale af denne kaliber vil være både tids- og ressourcekrævende.Machine learning teknikker samt regular expression “udtryksordbøger” er blevet anvendt til at hjælpe forskere med at udføre systematiske reviews, identificere fejl i citat gennemgang og tillade automatisk kategorisering af videnskabelige studier. Disse værktøjer er frit tilgængelige online, med henblik på at fremtidige forskere kan anvende- og videreudvikle dem, til gavn for fremtidig forskning i dette felt. Disse værktøjer hjælper med at evaluere hvilke medicinske præparater, der har vist sig effektive og under hvilke omstændigheder, samt kvalitet af dette data. Der er blevet foretaget to særskilte analyser for at belyse:

1) anvendelsen af ketamin som effektiv antidepressiv stof og 2) anvendelsen af dyremodeller af depression med ændret sammensætning af tarmbakterier. 3) På grundlag af den systematisk review af ændret sammensætning af tarmbakterier blevet der udført et eksperimentel dyreforsøg for at øge forståelsen af effekten af præbiotika i en genetisk dyremodel (Flinders Sensitive Line rotter).

Disse undersøgelser viste: 1) Samlet kvantitative data viste at ketamin er en effektiv antidepressiv præparat. 2) Samlet kvantative data viste at mange forskellige slags indgreb er blevet fortaget for at ændre sammensætning af tarmbakterier hvor, nogle virker som antidepressiv midler og andre som depressiv midler. Der var ingen forsøg som undersøgt præbiotika. 3) Dette forsøgte jeg at undersøge ved at tilføre ikke-fordøjelige kostfibre til rotters almindelig kost og undersøg effekten på depressiv- og angst-opførsel i en genetisk dyremodel på Flinders Sensitive Line rotter.

De nyudviklede værktøjer har forbedret vores overblik over anvendte dyremodeller i depression samt virkningen af antidepressiv medicin. Resultere af systematisk review er med til at skabe et overblik over tilgængelig forskning om depression: a) hvilken effekt forskellige præparater har i forsøg med dyremodeller, b) hvordan forskellige faktorer, såsom stress og genetiske ændringer, påvirker udviklingen af depression. Dette kan give indblik i c) de underliggende biologiske mekanismer bag behandling og eksponering. Disse analyser forbedrer beslutningsgrundlaget for udvælgelse af mere målrettet og resultatskabende forskning. Dyreforsøgene kan blive tilrettelagt mere specifikt og viden om virkning af testede præparater er øget. I fremtiden kan disse værktøjer bruges på mange andre forskningsområder, således at forskningsmidler kan anvendes mere målrettet og ny viden kan sammenlignes med tidligere resultater, hvilket kan gavne patienter, behandlere og samfundet.

Declaration

I declare that the thesis has been composed by myself and that the work has not be submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where explicitly stated otherwise by reference or acknowledgment in the text, or where work which has formed part of jointly-authored publications has been included. My contribution and the contribution of the other authors to this work have been explicitly indicated in each chapter. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others. The work presented in Chapter 2 was previously published in Evidence-Based Medicine as “Understanding in vivo modelling of depression in non-human animals: a systematic review protocol” by myself, Alexandra Bannach-Brown, Jing Liao, and my supervisors; Gregers Wegener, and Malcolm Macleod. This study was conceived by all of the authors. I developed the protocol and wrote the manuscript with supervision from Jing Liao, Gregers Wegener, and Malcolm Macleod.

An earlier published version of the work presented in Chapter 3 was published as a pre-print on BioRxiv as “The use of text-mining and machine learning algorithms in systematic reviews: reducing workload in preclinical biomedical sciences and reducing human screening error” by myself, Alexandra Bannach-Brown (ABB), Piotr Przybyła (PB), James Thomas (JT), Andrew S.C. Rice (AR), Sophia Ananiadou (SA), Jing Liao (JL), and Malcolm Robert Macleod (MRM). I screened and analysed the datasets. JT & PB conducted feature selection and built the classifiers. ABB, JT & PB wrote the manuscript. ABB, JT, PB, MRM, JL, AR & SA devised the study. JL, MRM & SA supervised the study.

The protocol for the animal experiment in Chapter 7 has been accepted for publication in BMJ Open Science. The experiment in Chapter 7 was designed with help from Sandra Tillmann, a PhD student at the Translational Neuropsychiatry Unit, Aarhus University and Professor Gregers Wegener. Professor Malcolm Macleod advised on the *a priori* sample size calculation and statistical analysis plan.

Signed: Alexandra Bannach-Brown

_____

AARHUS UNIVERSITY

Declaration of co-authorship concerning article for PhD dissertations

Full name of the PhD student: Alexandra Bannach-Brown

This declaration concerns the following article/manuscript:

Title:	Understanding in vivo modelling of depression in non-human animals: a systematic review protocol
Authors:	Bannach-Brown, A., Liao, J., Wegener, G., & Macleod, M.R.

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference: Bannach-Brown, A., Liao, J., Wegener, G. & Macleod, M., 2017, "Understanding in vivo modelling of depression in non-human animals: a systematic review protocol". Evidence-based Preclinical Medicine. 3, 2, p. 20-27.

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

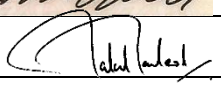
The PhD student has contributed to the elements of this article/manuscript as follows:

A. Has essentially done all the work
B. Has done most of the work (67-90 %)
C. Has contributed considerably (34-66 %)

D. Has contributed (10-33 %)
E. No or little contribution
F. N/A

Element	Extent (A-F)
1. Formulation/identification of the scientific problem	A
2. Development of the method	A
3. Planning of the experiments and methodology design and development	A
4. Involvement in the experimental work/clinical studies/data collection/obtaining access to data	F
5. Development of analysis plan and preparation of data for analysis	A
6. Planning and conducting the analysis of data	A
7. Interpretation of the results	F
8. Writing of the first draft of the manuscript	A
9. Finalization of the manuscript and submission	A

Signatures of first- and last author, and main supervisor

Date	Name	Signature
18/03/2019	Alexandra Bannach-Brown	
18/3/19	Malcolm Macleod	
18/3/19	Gregers Wegener	

Date: 18/03/2019

Signature of the PhD student

AARHUS UNIVERSITY

Declaration of co-authorship concerning article for PhD dissertations

Full name of the PhD student: Alexandra Bannach-Brown

This declaration concerns the following article/manuscript:

Title:	Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error
Authors:	Alexandra Bannach-Brown, Piotr Przybyła, James Thomas, Andrew S. C. Rice, Sophia Ananiadou, Jing Liao, Malcolm Robert Macleod

The article/manuscript is: Published ☒ Accepted ☐ Submitted ☐ In preparation ☐

If published, state full reference: Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A.S., Ananiadou, S., Liao, J. and Macleod, M.R., 2019. Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic reviews*, *8*(1), p.23.

If accepted or submitted, state journal:

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

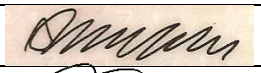
The PhD student has contributed to the elements of this article/manuscript as follows:

A. Has essentially done all the work

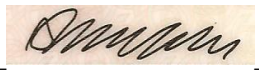
B. Has done most of the work (67-90 %)
C. Has contributed considerably (34-66 %)
D. Has contributed (10-33 %)
E. No or little contribution
F. N/A

Element	Extent (A-F)
1. Formulation/identification of the scientific problem	B
2. Development of the method	E
3. Planning of the experiments and methodology design and development	C
4. Involvement in the experimental work/clinical studies/data collection/obtaining access to data	B
5. Development of analysis plan and preparation of data for analysis	C
6. Planning and conducting the analysis of data	B
7. Interpretation of the results	B
8. Writing of the first draft of the manuscript	B
9. Finalization of the manuscript and submission	B

Signatures of first- and last author, and main supervisor

Date	Name	Signature
18/03/2019	Alexandra Bannach-Brown	
18/3/19	Malcolm Macleod	

Date: 18/03/2019



Signature of the PhD student

AARHUS UNIVERSITY

Declaration of co-authorship concerning article for PhD dissertations

Full name of the PhD student: Alexandra Bannach-Brown

This declaration concerns the following article/manuscript:

Title:	Administration of galacto-oligosaccharide prebiotics in the Flinders Sensitive Line animal model of depression
Authors:	Alexandra Bannach-Brown Sandra Tillmann Malcolm Robert Macleod Gregers Wegener

The article/manuscript is: Published ☐ Accepted ☒ Submitted ☐ In preparation ☐

If published, state full reference:

If accepted or submitted, state journal: BMJ Open Science

Has the article/manuscript previously been used in other PhD or doctoral dissertations?

No ☒ Yes ☐ If yes, give details:

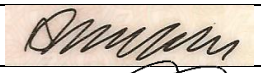
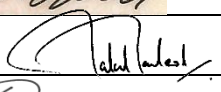
The PhD student has contributed to the elements of this article/manuscript as follows:

G. Has essentially done all the work

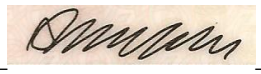
H. Has done most of the work (67-90 %)
I. Has contributed considerably (34-66 %)
J. Has contributed (10-33 %)
K. No or little contribution
L. N/A

Element	Extent (A-F)
1. Formulation/identification of the scientific problem	A
2. Development of the method	C
3. Planning of the experiments and methodology design and development	B
4. Involvement in the experimental work/clinical studies/data collection/obtaining access to data	B
5. Development of analysis plan and preparation of data for analysis	B
6. Planning and conducting the analysis of data	A
7. Interpretation of the results	A
8. Writing of the first draft of the manuscript	A
9. Finalization of the manuscript and submission	B

Signatures of first- and last author, and main supervisor

Date	Name	Signature
18/03/2019	Alexandra Bannach-Brown	
18/3/19	Malcolm Macleod	
18/3/19	Gregers Wegener	

Date: 18/03/2019



_____ _____

Signature of the PhD student

Acknowledgements

Firstly, I'd like to thank my grandparents and parents, Lani and Douglas, for imparting the importance of education, for encouraging me and facilitating me to pursue this field.

I am grateful for this unique opportunity to conduct research between Aarhus University and the University of Edinburgh, and to learn from inspiring and ground-breaking researchers, above all Prof. Malcolm Macleod. This unique collaboration has opened up several opportunities that I have been fortunate to be able to pursue. Firstly, being awarded a fellowship from ContentMine to develop text-mining tools for evidence synthesis. Secondly, working with the Systematic Living Information Machine collaboration to develop machine learning algorithms has been a fascinating experience into the world of automation which I will continue to dive deeper into. I have been inspired by shadowing Professor Andrew McIntosh in his psychiatry clinic and working with NHS Lothian Child & Adolescent Mental Health Services as an assistant psychologist. I am grateful for the opportunity to teach statistics and experimental design, at the University of Edinburgh and Aarhus University, as well as supervising many brilliant students. I'd like to acknowledge the work of Anthony Shek, Grace Wallace, Fraser Sneden, and Oskar Jefsen, who contributed to research in this thesis.

Thank you to Malcolm and Dr Emily Sena for creating the wonderful team that is CAMARADES Edinburgh. I am indebted to you all for teaching me systematic review and meta-analysis, for all the sense-checks, and the ethos of open research and collaboration that will stick with me throughout my career. Thank you to Dr Sarah McCann for keeping me right, Dr Jing Liao for your patience and helping me find bugs in my code, and to my PhD partner, Dr Zsanett Bahor, here's to many more academic adventures. Cheers all, happy Gin Friday! Thank you to Prof. Gregers Wegener for welcoming me into your lab, I'd like to acknowledge you and the team at TNU for the insight into performing animal work and your assistance in designing my experiments, particularly Dr Sandra Tillmann and the office gang. I am grateful for the support of friends throughout, particularly The Cali Girls and Dr Nina Fisher, thank you for always being there.

Thank you to my PhD examiners, Prof. David Cunningham Owens, Prof. Iain Marhsall, Prof. Catherine Belzung, and Prof. Søren Østergaard. I have greatly appreciated your expert contribution and the valuable academic discussion during the closed viva, which I believe has improved the quality of the thesis.

An investment in knowledge pays the best interest. – Tak Kurt & Britta

Table of Contents

1 INTRODUCTION

1.1 Problem Statement

Depression is the leading cause of disability worldwide impacting an estimated 322 million people worldwide with a global prevalence of 4.4% (World Health Organisation, 2017). Current treatments are adequate at best. Many patients are non-responsive to the current treatments and many patients do not receive the treatment they require. A number of challenges exist which decrease the speed of discovering new viable treatments. Most importantly, the pathophysiological mechanisms underlying this complex disorder are not yet understood, there is large heterogeneity in the symptoms present across patients. Drug discovery in psychiatry is slow. Given the high experimental control of animal models, they have been used to elucidate the underlying mechanisms of antidepressant treatments and pathophysiology. Systematic review is a method of openly and systematically collecting and synthesising the available evidence of a given topic. Systematic review is a hypothesis-generating tool and meta-analysis can be carried out subsequently, by statistically pooling the data to produce an estimate of overall effect.

The aim of this thesis is to apply systematic review and meta-analysis techniques to synthesise the evidence available on animal models of depression and antidepressants investigated. Further, to achieve an overview of what species are used, and what techniques have been used to induce depressive-like phenotypes. We can get an understanding of which treatments have been investigated in animal models of depression, how confident can we be in these findings, and what information is required to translate these findings accurately to useful treatments for the clinical population? Findings from systematic review and meta-analysis can inform and refine the design of primary animal experiments, however, they are resource-consuming tools. Therefore, to facilitate the research process of evidence synthesis, automation tools have been developed, tested, and implemented to provide an overview of the literature more quickly.

In this thesis I provide an overview of the literature on animal models of depression to inform primary *in vivo* research. I apply automation tools to this systematic review of animal models of depression. Two systematic reviews of the available literature on

ketamine and microbiota-targeting interventions in animal models of depression were conducted. Findings from the systematic review of microbiota-targeting interventions were used to inform a primary animal experiment on prebiotics in Flinders Sensitive Line animals. We can use findings from a systematic review to better understand factors that impact the efficacy of treatments in animal models and improve the quality of research conducted, to better understand depression and improve the treatments available to patients.

1.2 Depression

1.2.1 Prevalence and burden of depression

Depression is the leading cause of disability in the world (Marcus et al., 2012) and is currently the brain disorder with the highest financial cost in Europe (Gustavsson et al., 2011). The number of people diagnosed with depression worldwide is estimated at 322 million with a global prevalence of 4.4% (WHO, 2017). Major Depressive Disorder (MDD) is a mental illness characterised by "*low mood, loss of interest and pleasure or loss of energy*" (DSM-5 & ICD-10). The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) and the International Statistical Classification of Diseases and Related Health Problems (ICD-10) guidelines differ slightly as to the core symptoms required for a diagnosis of depression. The DSM-5 requires that patients present with 5 out of the 9 symptoms for at least 2 weeks, where at least one symptom being depressed mood or loss of interest. The ICD-10 requires patients to present with two out of the first three symptoms; depressed mood, loss of interest, and reduction in energy, for at least two weeks, with two subsequent symptoms from the remaining symptoms (NICE Guidelines, 2010), see table 1.1.

Table 1.1 An overview of symptoms required for a diagnosis of depression in the ICD-10 and DSM-5 diagnostic manuals. The asterisk denotes the core symptoms required for diagnosis.

ICD-10	**DSM-5 Major/Minor Depression**
Depressed Mood*	Depressed mood by self-report or observation made by others*
Loss of interest*	Markedly diminished interest or pleasure*
Reduction in energy*	Fatigue/loss of energy
Loss of confidence or self-esteem	Feelings of worthlessness/excessive or inappropriate guilt
Unreasonable feelings of self-reproach or inappropriate guilt	Recurrent thoughts of death, suicidal thoughts or actual suicide attempts
Recurrent thoughts of death or suicide	Diminished ability to think/concentrate or indecisiveness
Diminished ability to think/concentrate or indecisiveness	Psychomotor agitation or retardation
Change in psychomotor activity with agitation or retardation	Insomnia/hypersomnia
Sleep disturbance	Significant weight loss/gain or increased/decreased appetite
Change in appetite with weight change	

Depression places a huge burden on patients and is a great cost to healthcare systems and governments. It is estimated to contribute over 50 million Years Lived with Disability globally (WHO, 2017). Mood disorders are the largest contributors to the European cost of brain disorders, costing €113.4 million out of the total €476.9

million direct healthcare costs (Gustavsson et al., 2011). Depression reduces life expectancy by an average of 7-11 years (Chesney, Goodwin & Fazal, 2014).

The severity of a major depressive episode and associated functional impairment is part of the diagnosis and is classified into three broad classes by the ICD-10 and the DSM-5; mild, moderate and severe. These classifications are made based on the number of symptoms patients present with, with less (4) symptoms required for a diagnosis of depression in the ICD-10 (see table 1.2). This approach, the counting of symptoms, is in line with the idea that depression is a consistent syndrome (Fried & Nesse, 2015a). However, the utility of this approach has been called into question as it does not take into account individual symptom severity or the level at which the individual is functionally impaired by the disorder (NICE Guidelines, 2010). Clinicians are advised to assess individual patients on their degree of functional impairment before making a diagnosis (NICE Clinical Guidance, 2010). Further, symptoms of depression can be classified into indistinct subtypes; melancholia, catatonia, seasonal affective disorder, post-partum depression, and depression with psychotic features (psychotic depression). Classification of symptoms into subtypes may have an impact on the treatment intervention plan in certain cases, such as light therapy to seasonal affective disorder (NICE Guidelines, 2010). However these subtypes are critiqued for their validity and poor utility when implemented (Lichtenberg & Belmaker, 2010; Baumeister & Parker, 2012; Rush, 2007).

Table 1.2 Classification of major depression. Numbers refer to the number of symptoms required for each classification

	DSM-5 major depression	**ICD-10 depressive episode**
Mild	Minimal above the minimum (5)	4
Moderate	Between mild and severe	5-6
Severe	Several symptoms in excess of 5	7+

Depression is a disorder that persists across the lifetime, often with relapsing and remitting periods. Furthermore, symptoms can persist between episodes (NICE

Clinical Guidance, 2009). The primary aim of interventions is complete relief of symptoms (NICE, 2009). Depressive episodes can last weeks to years, and the probability of recovery from a depressive episode decreases the longer an episode lasts (Pattern, 2006). The probability of experiencing another depressive episode is greatly increased after experiencing an initial episode. Throughout this thesis when I refer to “depression”, I mean the range of symptoms which are characteristic of the clinical disorder, major depression disorder.

There are currently several interventions implemented as part of the National Health Service (NHS) in the United Kingdom to reduce the symptoms of depression. These interventions include exercise, psychological therapies, and pharmacological interventions, or a combination of these, depending on the severity of the depression experienced (NHS, 2014).

The rate of remission with antidepressant medication is at best 70%, and may only be achieved after several different interventions have been tried (Rush et al., 2006; Geddes et al, 2003). Psychotherapy has comparable efficacy as pharmacological interventions in patients with mild to moderate MDD (Spielmans et al., 2011). Currently, treatment-resistance is highly prevalent among patients. This is estimated to be approximately 50% of patients that have moderate to severe depressive symptoms and do not respond to more than 4 “lines” of pharmacological treatments (Rush et al., 2011).

1.2.2 Risk factors

Depression commonly manifests in adolescence or early adulthood, between the ages of 14 and 24, although can occur earlier or later in life (Kessler et al., 2005). There are several risk factors that are known to be associated with the increased risk of developing depression as well as the progression and recurrence of the disease across the lifetime. These interact to influence the risk of depression. These include a range of environmental, sociodemographic, and genetic factors (Colman et al., 2014; Kessler et al., 2005), such as being female, acute and chronic stressful life events in childhood such as separation from primary caregiver or parental divorce, and

adulthood such as living in poverty, death of a loved one or an employment crisis, low birth weight and delayed age of reaching childhood developmental milestones (Colman et al., 2014; Kessler et al., 2003). A recent genome-wide association study identified 44 single nucleotide polymorphisms associated with an increased risk of experiencing MDD (Wray et al., 2018).

1.2.3 Heterogeneity

Depression presents heterogeneously, with many adjunctive symptoms and sub-types or categories. The possible combination of symptoms that are needed to classify for a clinical diagnosis contributes to this heterogeneity. At least five out of nine symptoms under DSM-5 are required for a clinical diagnosis, where several key symptoms can present with dimensionality e.g. insomnia/hypersomnia, weight loss or gain. Under this classification system, two patients can be diagnosed with MDD but not share symptoms in common, as shown in the table highlighted by Dzirasa and Convington (2012) (see table 1.3 below).

Fried and colleagues note that the DSM criteria of depression can *"lead to 1497 potential unique symptom profiles that all qualify for the same diagnosis (Ostergaard et al., 2011), including profiles that do not have a single symptom in common"* (Fried et al., 2014). Taking into account the sub-types and dimensionality of some of the symptoms, e.g. sleep disturbance which may feature hyper- or hyposomnia, increases the number of possible unique profiles to 16,400 (Fried & Nesse, 2015b), remembering that each symptom contributes equally to the diagnosis (Fried et al., 2014). Fried and Nesse analysed the presence of symptom profiles in a well-known dataset, the STAR*D cohort, and identified 1,030 unique profiles across the cohort of 3,703 patients. 24% of patients described the most common 30 symptom profiles, and 13.5% of patients had completely unique symptom profiles. The most common symptom profile was only prevalent in about 2% of patients. This highlights the huge variability in the presentation of depression. This finding may also have implications for the current approach in psychiatry for diagnosis, mainly the approach of using symptom counting to assess depression severity, which is critiqued for being inadequate for gaining full insights and understanding into the condition from clinical symptoms (Fried & Nesse, 2015b).

This heterogeneity in the human condition can impact the apparent efficacy of treatments tested in a clinical trial, in that two patients with a diagnosis of a depression could present with two completely different symptom profiles without overlap of a single symptom. When trying to study depression and the efficacy of treatments in clinical trials, the symptom profile of a patient is largely ignored during recruitment and the current approach of symptom counting is favoured.

Table 1.3. Symptom presentation for two hypothetical patients (Dzirasa & Convington, 2012).

Patient 1	***Patient 2***
Poor Mood	Anhedonia
Hypersomnia	Insomnia
Feelings of worthlessness	Decreased concentration
Psychomotor retardation	Significant weight loss
Fatigue	Suicidal thoughts

To further muddy the waters, inter-rater reliability between psychiatrists' diagnosis of MDD can be poor. In the DSM-5 field tests, Reiger and colleagues captured the lower end of this spectrum, reporting kappa values of 0.25 between psychiatrists' diagnosis of major depressive disorder (Reiger et al., 2013). This can present difficulties in 'characterising' groups of patients with depression for the testing of treatments in clinical trials, unpicking why some patients respond and others do not. The poor agreement between clinicians in the diagnosis of symptoms can add further complexity when preclinical evidence for treatments fail to translate to clinical trials, as is seen in many other preclinical models (O'Collins et al., 2006).

To try and understand why there is such a wide variety in symptom profiles and if symptom profiles can help target treatment strategies, researchers have attempted to

cluster symptoms using a data-driven approach from studies of clinical populations. This aims to empirically create sub-types of depression that are more informative than the current subtype classification. A central theoretical rationale behind this approach to understanding the heterogeneity in depression is that patients, in fact, suffer from different syndromes with different aetiology, predisposing factors, and symptom presentation (Fried & Nesse, 2015a).

Various techniques, such as factor analysis, principal components analysis and latent cluster analysis, are used to identify patterns of symptoms across large groups of patients with depression, and to identify latent variables based on similarity in the data from patients. These techniques have been applied to a number of datasets with different cohorts of patients who have been diagnosed with the help of different tools. A systematic review of these approaches was conducted by van Loo and colleagues (2012). Although there are a number of findings that support a finding of cognitive or somatic symptom dimensions, there is inconclusive evidence to suggest that the finding of symptomatic subtypes hold true across methods and across datasets. The results vary based on the clustering technique used, which screening tool was used, and across patient samples. These differing results from symptom-cluster studies have lead other researchers in the direction of using network models to understand heterogeneity in symptoms. This enables an understanding of the direct and indirect influences on dynamic causal networks connecting symptoms and investigating which symptoms “cluster” (Fried, 2015). Findings from these new network models may provide important information for understanding the link between symptoms and may provide insight into underlying mechanisms (Fried, 2017).

Despite decades of investigation into depression, little is known about the biological mechanisms underpinning the disease (Nestler, Gould & Manji, 2002; Slattery & Cryan, 2014). With a better understanding of the mechanisms causing the development and progression of depression, the development of novel and more reliable treatments might be possible. There is solid rationale that further investigation into the mechanisms and factors that contribute to the development of depression is needed. This is a highly important area to tackle, both from a clinical, and a preclinical perspective. The National Institute of Mental Health in the USA has recognised that the current diagnostic approach lacks validity (Insel, 2013), which explains their recent

decision to only fund projects that aim to elucidate underlying neurobiological mechanisms. Others also call for a novel approach to diagnosis under a slightly different methodology. One such approach is the concept of 'Symptomics', the investigation of psychopathology at the level of individual symptoms, aiming to understand associations between symptoms to try and untangle cause and effect relationships (Fried, 2017; Fried, 2015)

This section has introduced the prevalence and burden of depression, the symptoms of depression, the subtypes of depression, risk factors, and the current method used for diagnosis in clinical practice. The heterogeneity of the disorder has been discussed and the next section will build on this to highlight the limitations of the current paradigm.

1.3 Treatments for MDD

1.3.1 Current Pharmacological Interventions

Historically the discovery of modern pharmacological drugs occurred by serendipity in the early 1950s in the search for pharmacological treatments for tuberculosis (Hillhouse & Porter, 2015). Trials of iproniazid for the treatment of tuberculosis produced "side-effects" such as psychostimulation, increased appetite, and improved sleep, which were then tested formally in patients with depression (Loomer, Saunders & Kline, 1958) which showed a 70% improvement in symptoms. Isoniazid was the first monoamine oxidase inhibitor (MAOI). MAOIs act by inhibiting monoamine oxidase A and monoamine oxidase B, enzymes throughout the body that catalyse the oxidation monoamine neurotransmitters, thus reducing the breakdown of key neurotransmitters such as serotonin, norepinephrine, and dopamine, increasing the availability of these neurotransmitters around the synapse (Youdim, Edmondson, & Tipton, 2006). Iproniazid was marketed as an antitubercular compound but used off-label for MDD (Hillhouse & Porter, 2005).

The initial non-selective MAOIs, were discontinued shortly after introduction due to adverse side-effects and serious concerns regarding interactions with foods containing large amounts tyramine such as cheese, causing increased heart rate and sweating (Hillhouse & Porter, 2015). Attempts were swiftly made to alter the

pharmacodynamics of early MAOIs to make them more selective and reversible. Currently marketed MAOIs include; selegiline and moclobemide. Rates of efficacy for modern MAOIs are comparable to other traditional antidepressants, with reports between 50-70% of response rate in patients (Krishnan, 2007; Henkel et al., 2006). Common side effects reported include insomnia, gastrointestinal disturbances, and some sexual dysfunction (Shulman, Herrmann & Walker, 2013).

Concurrently, compounds that had antipsychotic properties for schizophrenia were being tested, one of which, imipramine, failed to have an antipsychotic impact but did reduce symptoms in depressive patients. Imipramine, modified from the antihistamine structure, was given FDA approval for major depressive disorder and thus the advent of tricyclic antidepressants (TCAs) began. TCAs have been shown to act via a broad range of mechanisms but mainly via the serotonin transporter and the norepinephrine transporter. Further, many TCAs act as histamine H1 receptor antagonists. Other mechanisms include selectively blocking serotonergic, adrenergic and muscarinic receptors (Tatsumi et al., 1997; Hillhouse & Porter, 2015).

Currently marketed TCA drugs include; imipramine, desipramine, amitriptyline, and clomipramine (Gillman, 2007). TCAs perform well in comparison to other antidepressants. In a recent meta-review, TCAs had a medium effect size (relative risk of response to treatment) with moderate strength of evidence (Gartlehner et al., 2016), similar to previous reports and to the efficacy of MAOIs (Krishnan, 2007). A more recent network meta-analysis found amitriptyline to be the most effective antidepressant among 21 licensed drugs (Cipriani et al., 2018). Despite their efficacy, TCAs have a number of serious side-effects including cardiotoxicity, cognitive and psychomotor dysfunction (Montgomery & Kasper, 1995). Research around the mechanisms of TCAs and pharmacological work to reduce adverse side-effects lead to further discovery of groups of drugs known as selective serotonin reuptake inhibitors (SSRIs), serotonin–norepinephrine reuptake inhibitors (SNRIs) and norepinephrine reuptake inhibitors (NRIs). Research into TCAs, together with the discovery of the monoamine oxidase inhibition by iproniazid and the noradrenaline reuptake inhibition of imipramine, fuelled the postulation of the Catecholaminergic Hypothesis by Schildkraut in 1965 who stipulated that depression was the result of a

reduction of noradrenaline in and around the synaptic cleft (López-Muñoz & Alamo, 2009).

Serotonin reuptake inhibitors (SSRIs) were the first drug to be discovered through rational drug design. Rational drug design involves identifying a target receptor or enzyme associated with the disease, the target receptor is characterised, and a molecule that targets that receptor, whilst ideally not binding to or targeting similar receptors or enzymes, to reduce side-effects (Todd, Anderson & Groundwater, 2009). The exact mechanisms of SSRIs have not yet been fully elucidated, but the primary mechanism is thought to be through blocking the uptake of serotonin at the presynaptic neuron. SSRIs have also been shown to possess a low binding affinity for some postsynaptic receptors, for example, dopamine D2 and adrenergic alpha-1, alpha-2 and beta (Owens et al., 1997). Fluoxetine was the first marketed SSRI in 1987 (López-Muñoz & Alamo, 2009). This discovery sparked the study of theories surrounding the role of neurotransmission in depression. Following on from the Monoamine Hypothesis, the Serotonin Hypothesis was proposed, which postulated that a deficit of serotonin in the brain was the cause of depression (Albert, Benkelfat & Descarries, 2012). The early evidence for this theory came from studies of post-mortem brain samples from suicide patients which showed decreased serotonin (Shaw, Camps & Eccleston, 1967). Currently marketed SSRI drugs include; citalopram, fluoxetine, escitalopram, fluvoxamine. These SSRIs perform better than placebo, overall have medium acceptability and have fewer dropouts due to adverse side-effects (Cipriani et al., 2009; Cipriani et al., 2018).

The discovery of fluoxetine and the rise of the Serotonin Hypothesis of Depression sparked the further investigation into serotonin modulation in the central nervous system (López-Muñoz & Alamo, 2009). With this came a wave of new classes of drugs named after their selective or combined modulation, stimulation, antagonism or reuptake inhibition of serotonin and norepinephrine; Norepinephrine Reuptake Inhibitors (NRIs), Serotonin & Norepinephrine Reuptake inhibitors (SNRIs), Serotonin Modulators & Stimulators (SMS), Serotonin Antagonists & Reuptake Inhibitors (SARIs), and Norepinephrine–Dopamine Reuptake Inhibitors (NDRIs). The new wave of drugs was classified by their main mechanism of action. For example, dual serotonin and norepinephrine reuptake inhibitors (SNRIs) inhibit the reuptake of

serotonin and norepinephrine at the synapse to increase the concentration of these neurotransmitters at the synaptic cleft. Commonly prescribed drugs include; Bupropion (NDRI), Reboxetine (NRI), Trazodone (SARI), Vilazodone and Vortioxetine (SMS), Venlafaxine, Duloxetine, Levomilnacipran, and Milnacipran (SNRIs). These drugs perform similarly to SSRIs, they generally perform moderately better than placebo and have good acceptability profiles (acceptable drop-out rates or discontinuation due to adverse side-effects).

To summarise this section, a range of antidepressants are currently available, sparked out of serendipity in the early 1950s. Current antidepressants are more efficacious than placebo with modest effect sizes (Cipriani et al., 2018; Cipriani et al., 2009). Drop-outs due to side-effects, remission, and mood symptoms were investigated for 21 different pharmacological treatments. Agomelatine, amitriptyline, escitalopram, mirtazapine, paroxetine, venlafaxine, and vortioxetine were found to be most effective. Critique of the current approach of classifying antidepressants will be discussed below.

1.3.2. Classification of Antidepressants

The current classification of antidepressant drugs has been critiqued for a number of reasons. Firstly, many antidepressants act through several mechanisms, potentially targeting several neurotransmitter receptors both pre- and post-synaptically. Secondly, although meta-analyses show antidepressants to be moderately more effective than placebo, there is still a high incidence of treatment-resistant depression with reports up to 50% of patients not responding after 2 lines of pharmacological intervention (STAR*D trial, Huynh & McIntyre, 2008). Thirdly, it is common that pharmaceuticals originally marketed for depression, for example, are prescribed for patients with other psychiatric disorders such as schizophrenia and Attention-Deficit-Hyperactivity Disorder (ADHD). Therefore, a task force from the European College of Neuropsychopharmacology (ECNP) is attempting to reclassify the current antidepressants and create a new nomenclature for newly developed drugs.

The task force from ECNP is developing a new nomenclature for classifying the currently available neuropsychiatric drugs. This classification is based on the pharmacology and known modes of action on the neurotransmitter or system (NbM2, Zohar et al., 2014). Although not all of the modes of action, antagonism and agonism, are known for each drug, the rationale is that it will help patients, by presenting the most cutting-edge knowledge from neuroscience and psychiatry. The task force has additionally added information on four dimensions concerning; approval from regulatory bodies, efficacy and known prevalent or life-threatening side-effects based on “solid” clinical evidence, a brief summary of the clinical knowledge for treating patients, and lastly a summary of the knowledge of neurobiology from clinical and preclinical studies (Zohar et al., 2014). Drugs are classified based on major and minor receptor agonism and antagonism.

For example, the drug desipramine, TCA, acts on norepinephrine as a reuptake inhibitor (Figure 1.1). The application shows that desipramine is approved for Major Depressive Disorder, and based on major clinical evidence, has been shown to be effective. The application goes on to describe the side-effects, the practical notes for prescription, and the underlying neurobiology with evidence from clinical and preclinical studies.

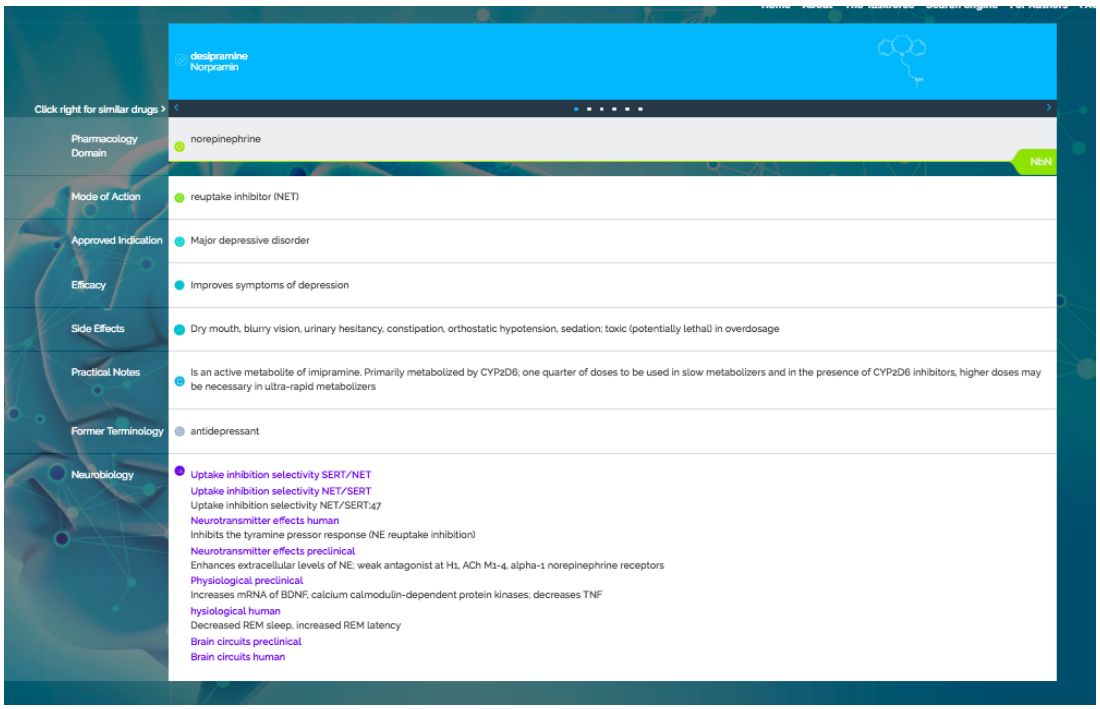


Figure 1.1. An example of an entry in the NbM2 nomenclature for psychiatric drugs, Despiramine.

One limitation of this approach is that it is based on the expertise of the members of the task force. Although the task force is comprised of world-leading expert psychiatrists and neuroscientists, the literature can be susceptible to publication bias, and it is not clear through what tools the literature has been searched for and collated, or whether this evidence will be quantified with meta-analysis of findings. With the application of automation tools for systematic evidence synthesis, many of these limitations can be eliminated. Even despite the current limitations, this approach shows great promise and could greatly impact the field of neuropsychiatry, closing the gap between clinical administration and current biological knowledge. The constant integration of new solid knowledge will ensure that patients receive advice on treatments from the most up to date knowledge.

1.3.3 Other Treatments for Depression

1.3.3.1 Other treatments for MDD

The above paragraphs outline the main pharmacological therapies for depression. Prior to the serendipitous discovery of pharmacological treatments, the main method of treatment was psychotherapy and psychoanalysis. Psychological therapies for depression have continued to develop alongside the development of antidepressants. The current primary therapies for depression include cognitive behavioural therapy (CBT), behavioural therapies, psychodynamic therapies, systematic therapies, and humanistic therapies. A comprehensive analysis of these therapies is beyond the scope of this work. Please refer to recent reviews and systematic reviews on the topic (for recent review of systematic reviews on psychological and pharmacological treatments see Gartlehner et al., 2016, for recent review on psychological therapies in depression see Shinohara et al., 2013; Linde et al., 2015a, Linde et al., 2015b). Other interventions for mild or sub-threshold depressive symptoms include exercise (NHS, 2016; Krogh et al., 2017) which generally decreases the risk of relapse, but the evidence synthesised is of poor quality, and prebiotics and probiotics (NHS, 2016). Other alternative therapies include light therapy in particular for seasonal affective disorder and dietary and herbal supplements such as omega-3 fatty acids and St. John's wort. Supplements show varying efficacy with insufficient strength of evidence (Gartlehner et al., 2016).

Comparatively, based on a review of systematic reviews, CBT has a medium effect size for reducing depressive-like symptoms. Dietary and herbal supplements, light therapy, and exercise, however, show varying efficacy with low or insufficient strength of evidence (Gartlehner et al., 2016).

1.3.3.2 Other treatments for treatment-resistant MDD

For more severe cases of depression and depression that does not respond to several lines of pharmaceutical interventions, electroconvulsive therapy (ECT) may be administered. ECT has been a treatment for major psychiatric conditions since 1938 (Cerletti & Bini, 1938 & Kalinowsky, 1939 in Abrams, 2002). The antidepressant effects of ECT have been shown to act through a number of mechanisms; inhibitory

neurotransmission and monoamine neurotransmitters, endocrinological pathways, and neurogenesis (Merkl, Heuser & Bajbouj, 2009). A systematic review and meta-analysis of clinical trials of ECT has shown its superior effect to placebo stimulation and slight superior efficacy in relation to pharmacotherapy for major depressive disorder (UK ECT Review Group, 2003). There are, however, a number of serious side effects including significant cognitive impact such as amnesia (Merkl, Heuser & Bajbouj, 2009). Despite its concerns, ECT is still used routinely, in particular for patients resistant to pharmacological interventions. Patients with severe depression and at risk of suicide report that the procedures have been life saving (Hersch, 2013).

More recently, ketamine, commonly used in anaesthesia, has been investigated as a rapid-acting, acute antidepressant and has been shown to be effective in patients with otherwise treatment-resistant depression (Williams & Schatzberg, 2016). It is thought to work through blocking Glutamatergic NMDA receptors. Animal studies have shown that ketamine upregulates BNDF and mTOR signalling, leading to synaptogenesis in prefrontal cortical neurons (Williams & Schatzberg, 2016). To date, Phase III clinical trials for ketamine in depression are still underway to assess the optimal route and number of administrations and to reduce possible psychotic side-effects (Trial: NCT02401139, Black Dog Institute Australia). Results from the completed clinical trials on ketamine show promising effects. Ketamine is markedly more effective than placebo at 24 hours, 3 days and 7 days post administration (McGirr et al., 2015). Ketamine shows promise to achieve FDA approval for clinical use and has the potential to reduce symptoms for many patients who do not achieve remission with standard pharmacological therapy.

1.4 Current Limitations in Clinical Research

Despite years of investigation into the underlying biological mechanisms and antidepressants, there are still a number of key issues that prevent us from fully understanding and ‘curing’ depression. Firstly, the efficacy of antidepressants and treatment resistance. Currently, treatment-resistance is highly prevalent among patients. Approximately 50% of patients with moderate to severe depressive symptoms do not respond to more than 4 different pharmacological interventions (Rush et al., 2011). The current treatments have reports of slow-acting symptom relief,

with up to 12 weeks before there is an onset of therapeutic action, which adds significantly to the disease burden (Murrough, 2012). There is an apparent discrepancy between the rising prevalence of depression and the reported efficacy of anti-depressant treatments which has been coined 'The Depression Conundrum' (Celie et al., 2017). Many large systematic reviews and meta-analyses including the recent network meta-analysis by Cipriani and colleagues (2018) have found that antidepressant pharmacological and psychological treatments are effective in comparison to placebo. With reports that prevalence rates of depression are on the rise, there is a discrepancy between these two pieces of information. This might be because there are problems delivering mental health and psychiatric services to patients, or patients having difficulties seeking and accessing help, or that the efficacy of treatments is overstated. Our understanding about treatment efficacy from large systematic reviews and meta-analyses, such as the recent network meta-analysis by Cipriani and colleagues (2018), has been critiqued in that publication bias is common and can affect the findings of efficacy from meta-analyses (Ioannidis, 2009). Therefore, it's key to understand what can differentiate patients that are likely to respond and those that are likely not to respond to certain treatments.

Secondly, the heterogeneity of the symptoms present in disorder has added to the complexity of understanding the underlying biological mechanisms and the search for novel treatments. The mainstay approach in psychiatry has been to understand depression as having a single cause and to search for the underlying mechanisms of symptoms of a latent disorder (Fried & Nesse 2015). As the above section has outlined, there is wide variation in the symptoms patients with depression present with, and there are multiple brain circuits and biological mechanisms involved in single symptoms (outlined below in section 1.5). What can be done to elucidate these patterns of symptoms and can they be useful for predicting treatment response or understanding the underlying pathological mechanisms?

A better understanding of the underlying biological mechanisms could assist in the search for reliable biomarkers for depression. Biomarkers that can reliably measure a disease state, a symptom, or a likely response to treatment and that is an inexpensive tool that can be widely implemented across healthcare settings, is ideal.

Some attempts to differentiate between patients that respond and do not respond to drug treatments include the use of structural equation modelling of positron emission tomography scans (PET), using activation patterns of limbic-cortical connections (Seminowicz et al., 2004). However, it is clear that our understanding of the molecular mechanisms underlying depression and treatment response will need to be improved. This includes the development of novel pathophysiological models, and understanding of the complex heterogeneity (Murrough, 2012), enabling discovery of accurate biomarkers to improve the treatment available to patients suffering from depression.

1.5 Current Understanding of Underlying Pathology & Mechanisms

This section will outline the key psychopathological and neurobiological mechanisms that are proposed or known to be underlying depression and depressive-like phenotypes. This section will discuss data and evidence from both animal literature and human literature. The structure of this section is highlighted in figure 1.2, starting with the hippocampus, its projections to the prefrontal cortex (PFC), the amygdala, and the ventral tegmental area (VTA) and nucleus accumbens (NAc). Finally, the hypothalamic-pituitary-adrenal axis (HPA axis) is discussed.

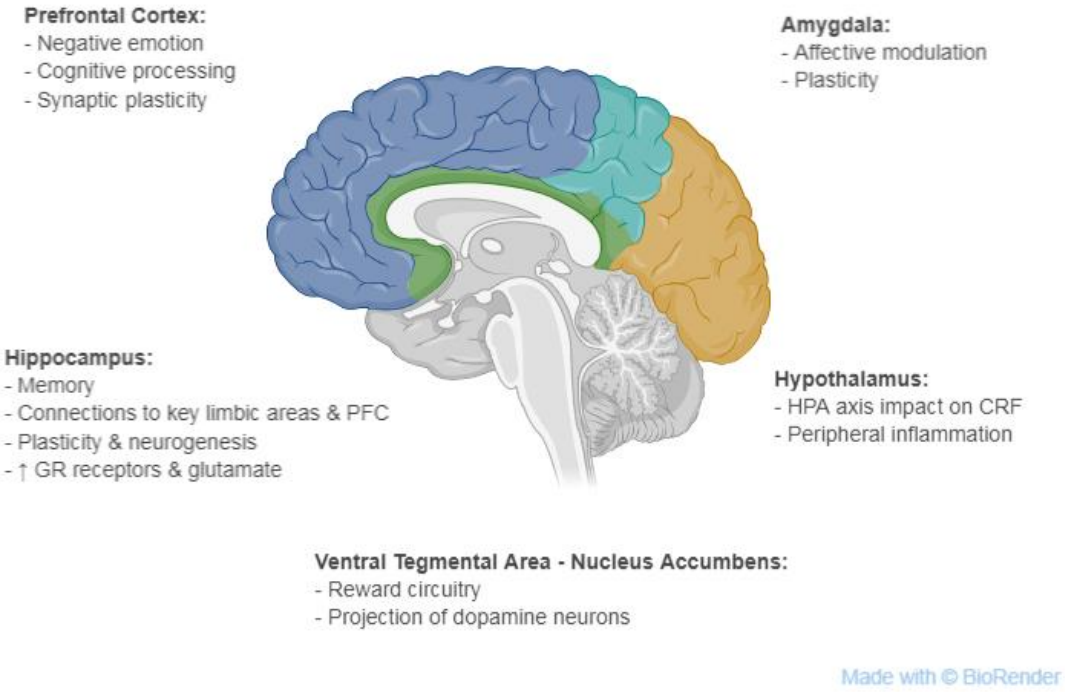


Figure 1.2. Main mechanisms involved in depression by brain area.

1.5.1 Hippocampus

The hippocampus is thought to be one of the most vital brain areas involved in depression due to its key role in several stress-related processes. The hippocampus is a key part of the limbic pathway and is connected to key brain areas involved in emotional processing, including the PFC and the amygdala.

The hippocampus is sensitive to stress which impacts plasticity, due to it being the primary site for adult neurogenesis (Lui et al., 2017). The key processes of synaptic transmission and connectivity are synaptic long-term potentiation and synaptic long-term depression. These are involved in memory formation and are affected after stress (Marsden, 2013). Regions of the hippocampus are affected differentially by stress and with different types of stress (Marsden, 2013). Stress decreases dendritic branching, the formation of new connections between neurons, decreases the generation of hippocampal neurons and increases cell death (Petrick et al., 2012). Depressed patients exhibit overgeneral processing of sensory information, with a

decrease in pattern separation, defined as being able to distinguish similar stimuli with non-overlapping neuronal representations (Belzung, Willner, Philippot, 2015). Depressed patients show an increase in pattern completion, which refers to being able to generalise stimuli in the case of partial sensory input (Belzung, Willner, Philippot, 2015). The granule cells in the dentate gyrus of the hippocampus are heavily involved in pattern separation (Yassa & Stark, 2011). With adult neurogenesis occurring mainly in the hippocampus, a decrease in adult neurogenesis with depression decreases pattern separation, with this association being bi-directional (Belzung, Willner, Philippot, 2015). Further converging evidence comes from the decreased volume of the hippocampus seen in patients with MDD, which is due to neuronal loss (Hanson et al., 2011). The primary hypothesis through which antidepressants are thought to be effective is through adult hippocampal neurogenesis. Many different antidepressants have been shown to induce hippocampal neurogenesis in humans and animals (Paizanis et al., 2007; Hamon & Blier, 2013), the processes is dependent on glucocorticoid receptor function (Anacker at al., 2011). The hippocampus has a high level of glucocorticoid receptors and glutamate which make it vulnerable to impact from HPA activation, which will be discussed further below.

1.5.2 Prefrontal Cortex

One of the main connections thought to be involved in depression is the connection between the hippocampus and the prefrontal cortex (PFC). Two main pathways are hippocampus-thalamus-PFC pathway, involved in attentional processing and memory, and the ventral hippocampus-basolateral amygdala-PFC involved in fear memory and social behaviour (Samptah, Sathyanesan & Newton, 2017). The PFC is a key centre for cognitive functioning, affection, attention, and goal-directed behaviour, which are implicated in depression and depressive-behaviour. The episodic buffer, a component of working memory integrating multimodal information, relies on the hippocampus and the PFC. The episodic buffer is involved in the development and maintenance of depressive schemata through the activation of depressive thought processes (Belzung, Willner, Philippot, 2015). Depressive thought processes are maintained by a feedback loop, increasing the bias toward the negative thoughts (Belzung, Willner, Philippot, 2015). Further, negative cognitive bias, the increased memory sensitivity for negative stimuli, is associated with increased activity in PFC, amygdala and

hippocampus (Belzung, Willner, Philippot, 2015). Rumination, is an important psychological process in depression. It describes the tendency to over analyse one's problems, feelings, and depressed mood states (Belzung, Willner, Philippot, 2015), and is found to correlate with decreased activity in the ventromedial PFC (Zhu et al., 2012). In depressed patients, different patterns of activity are involved, e.g. hyperactivity is seen in the ventromedial PFC and hypoactivity is seen in the dorsolateral PFC (Koenigs & Grafman, 2009).

Further, glutamate in the PFC plays a key role in depression. Ketamine, the NMDA-receptor antagonist, has been shown to increase the number of synapses and the synaptic function in the PFC through mammalian target of rapamycin (mTOR) and extracellular signal-regulated kinase (ERK) signalling pathways (Liu et al., 2017). Drugs that increase glutamate clearance prevent or reverse effects of chronic stress and chronic glucocorticoid exposure and exert antidepressant effects (Lang & Borgwardt, 2013) through synaptic plasticity. There is an association between immune activation and mechanisms involved in depressive symptoms, communication between the neuro-immune system and neural stem cells, which are able to regulate neuroplasticity (Krishnan & Nestler 2012; Eyre & Baune, 2012).

1.5.3 Amygdala

The hippocampus is further implicated in depression via its connection to the amygdala, which in turn is also closely linked to the PFC. The amygdala is an area of the brain strongly associated with emotional modulation, emotional memory encoding and fear conditioning in animal studies. However, dissimilar to the hippocampus and PFC, stress is found to induce dendritic branching, the formation of new connections between neurons (Marsden, 2013). Furthermore, brain-derived neurotrophic factor (BDNF) expression, a protein that encourages growth of new neurons and regulates of synaptic plasticity, is increased in areas of the amygdala under stress (Lang & Borgwardt, 2013). Several studies find a decrease in BDNF levels in the hippocampus after stress (Lee & Kim, 2010). BDNF levels are however decreased in plasma of patients with depression. The decrease in neurogenesis and BDNF may be mediated through the neuroimmune system, with evidence of a decrease of CD4+ cells leading to decreased neurogenesis (Wolf at al., 2009 in Eyre & Baune, 2012).

1.5.4 Ventral Tegmental Area & Projections

The hippocampus is further linked to the ventral tegmental area (VTA). The VTA is responsible for dopamine synthesis in neurons (Hamon & Blier, 2013). These dopamine neurons further project to nucleus accumbens and PFC (Oades & Halliday, 1987), part of the mesocorticolimbic pathway, and are involved in reward processing and aversion, as well as mediating short term to long term memory consolidation (Lui et al., 2017; Samptah, Sathyanesan & Newton, 2017). Increased firing rate of neurons in the VTA has been linked to susceptible and resilient animals after a repeated social defeat stress. This increase in firing rate is dependent on the projections of the neurons, with decrease firing rate seen in neurons projecting to the PFC (Chaudhury et al., 2013). However, stimulating VTA neurons using optogenetic techniques has also been shown to reverse the stress-induced deficits on the forced swim test and the sucrose preference test (Tye et al., 2013). The dopaminergic neurons from the VTA, PFC, and amygdala in the mesocorticolimbic pathway project to the nucleus accumbens (NAc), which is involved in the reward circuitry and motivation. Patients with depression show reduced activation in the NAc linked to the reward circuitry (Satterthwaite et al., 2015). The role of the VTA in depression and the complex interplay between stress and neurocircuitry will hopefully be further elucidated with more research (Polter & Kauer, 2015).

The lateral habenula, which receives input from the hippocampus and sends projections to the VTA has been implicated in depression and treatment. Deep brain stimulation (DBS) of this brain area has an antidepressant effect in patients with treatment-resistant depression (Sartorius et al., 2010). Studies in animal models of depression show that NMDA burst firing in the lateral habenula induced depressive-like behaviour (Yang et al., 2018). Infusions of ketamine into the lateral habenula cause rapid anti-depressant activity (Yang et al., 2018). This evidence suggests that the lateral habenula may play an import role in depression and anti-depressant mechanisms.

1.5.5 Hypothalamus

The hippocampus is further involved in depression through its sensitivity to stress toxicity through the overactivation of the HPA axis (Hamon & Blier, 2013). One of the major theories of depression is the cortisol hypothesis which suggests that an overactive HPA axis may mediate symptoms of depression via 1) the increased production of corticotrophin-releasing factor (CRF) from the hypothalamus, and 2) the reduced negative feedback at the central glucocorticoid receptors (Krishnan & Nestler, 2011). Glucocorticoid receptor functioning in the HPA axis is inhibited by proinflammatory cytokines such as interleukin-6 and tumour necrosis factor-alpha (Eyre & Baune, 2012). Increased HPA axis activity is seen in patients with depression. Higher levels of cortisol in the plasma and increased CRF in the cerebrospinal fluid are seen in MDD patients (Hamon & Blier, 2013; Mongeau et al., 2011). Antidepressants can restore cortisol levels by increasing the expression of glucocorticoid receptors which returns the feedback function to normal levels (Hamon & Blier, 2013). Glucocorticoid receptors can impact adult neurogenesis in the hippocampus. Overexpression of CRF during development in animal models leads to increased depressive-like behaviour in the forced swim test (Krishnan & Nestler, 2011). The HPA axis plays a pivotal role in depression through a number of mechanisms.

1.5.6 Other Mechanisms

In addition to neurobiological evidence, several other factors have been correlated with depression including metabolic disorders, such as obesity and diabetes. Leptin, a hormone that regulates energy balance in inhibiting hunger, acts at receptors in the hypothalamus. Leptin has been associated with mood symptoms as well as being involved in neurogenesis (Lawson et al., 2012). There appears to be a bi-directional risk between diabetes and depression, both diseases increase the risk of developing the other (Lang & Borgwardt, 2013). In addition, ghrelin, an amino-acid peptide involved in increasing hunger, is correlated with mood. Increased ghrelin is observed after acute and chronic stress (Chuang & Zigman, 2010) which may cause anti-depressant behaviour. Lastly, insulin growth factor increases hippocampal neurogenesis and can display anti-depressant behavioural responses (Lang & Borgwardt, 2013).

Non-neuronal correlates for depression have recently been investigated. Most prominently, the gut microbiota has been shown to play a role in depression, through the brain-gut-microbiota axis. This communication occurs via direct signalling through the vagus nerve and independently of the vagus nerve through complex metabolic, endocrine, immune and neural pathways (Dinan & Cryan, 2017). The evidence for the role of the brain-gut-microbiota axis will be further discussed in chapters 5 and 6.

To summarise, several brain areas and complex mechanisms are involved in depression including influences from peripheral mechanisms. Several streams of evidence converge to give us the current understanding of pathophysiological mechanisms behind the disorder but there are many questions still to be answered to provide the full picture. One way to help us elucidate the unknown mechanisms is by utilising animal models. The animal models currently used will be discussed in the section below.

1.6 Animal Models of Depression

Preclinical investigations contribute significantly to understanding the mechanisms underlying depression, which can, in turn, inform treatment development and have shown successful translation into clinical research. One example of this contribution is the investigations into the CLOCK gene's involvement in circadian rhythms and depression (Vitaterna et al., 1994; Bunney & Bunney, 2000). This research contributed to advances in successful treatments such as light therapy for depression and seasonal affective disorder (Eastman et al. 1998; Tuunainen et al., 2004). Preclinical experiments have the ability to model and dissect important mechanisms of action in the development and treatment of depression and can, therefore, provide insights into the neurobiology behind the disorder (Krishnan & Nestler, 2008). Additionally, preclinical experiments enable investigation into the safety and efficacy of proposed treatments prior to exposure in human cohorts (Kieburtz & Olanow, 2007), as well as the effect of treatments on a broad range of outcome measures, potentially characterising potential side effects. This knowledge can subsequently aid investigations into the optimal way to prevent the occurrence of depression, and the best and earliest interventions, which are top research priorities recently identified by

an MQ: Transforming Mental Health report (MQ: Transforming Health, 2016). Animal models contribute significantly to our understanding of depression.

1.6.1 Ethics of Animal Experiments

Animal experiments are a cornerstone of medical scientific discovery. Animal experiments have contributed significant understanding of underlying biological mechanisms and benefits to human health (Ioannidis, 2012; Comroe & Dripps, 1976).

However, frequently animal studies do not translate to benefits for human health due to poor quality and poor reporting of animal studies (Pound et al., 2004). Less than 10% of promising findings enter routine clinical use (Contopoulos-Ioannidis et al., 2003) and data from animal experiments was rarely considered during clinical trial design (Pound et al., 2004). The public acceptance of animal research and indeed the ethical implications of conducting animal research hinges on producing benefits for human (Pound et al., 2004). Organisations such as the National Centre for the Replacement Refinement and Reduction of Animals in Research in the UK, endorse and assist researchers and legislators to adhere to the principles of 3Rs (Replacement, Refinement & Reduction). The principles of 3Rs were introduced in 1959 (Russel & Burch, 1959) to reduce the inhumanity in animal experimentation. Replacement refers to methods used to avoid or replace the use of animals to answer scientific questions (Russel & Burch, 1959). Reduction refers to utilising appropriately designed methods to ensure that experiments are robust and reproducible, adding to the knowledge while using the minimum number of animals necessary and maximising the information gathered from each animal used (Russel & Burch, 1959). Refinement refers to methods to minimise the suffering and improve animal welfare (Russel & Burch, 1959). The NC3Rs in the United Kingdom work continuously to improve the techniques and methodologies used in animal experiments to ensure animal welfare (Prescott & Lidster, 2017).

Animal experiments are only informative for clinical trials if the results are valid and precise (Pound et al., 2004). The internal and external validity of animal experiments and the potential reasons for translational failure are discussed further in section 1.6.4.

There is sufficient evidence to suggest that animal experiments are not adequately reported. These findings come from systematic reviews of animal experiments summarising and pooling data. The utility of systematic reviews of animal experiments is outlined in Chapter 2.

Therefore, for animal ethics to be justified, we must ensure experiments are conducted accurately, that the maximal output is retrieved, that the minimum amount of suffering is gained whilst still getting the scientific benefit, and that we do get the benefit to human health (Pound & Bracken, 2014). If research is not conducted to a high quality, if the internal validity and external validity of the experiment are not considered and controlled, then the animal research may be considered unethical and not justified.

The following sections will discuss the ways in which animal models of depression are characterised. Additionally, it will discuss ways in which animal models are assessed, generally on two broad concepts; internal and external validity. Internal validity is defined as "*differences observed between groups of animals allocated to different interventions may, apart from random error, be attributed to the treatment under investigation*" (van der Worp et al., 2010, pg. 2). External validity refers to the generalizability of a study to a broader population beyond the animals investigated in the study, both to the same and to other species. The commonly used animal models in the literature will be discussed.

1.6.2 Characteristics of Animal Models of Depression

Broadly speaking, there are two types of interventions that are employed in the research on animal models. We distinguish between (1) interventions used to induce a model of depression or to mimic a depressive-like phenotype, and (2) interventions used to "rescue" depressive-like phenotypes through the testing of potential antidepressants. Further, we distinguish these from the measurement or assessment of depressive-like outcomes or variables of interest that could elucidate the underlying mechanisms behind the treatment effect or the effect of the 'model induction'.

In animal models of depression, the intervention being investigated is not always easy to distinguish. A model can be defined as including both a dependant variable (an outcome assessment/measurement) as well as an independent variable (model induction or manipulation) (Geyer & Markou, 1995). In the study of animal models of depression, the independent variable, “model induction”, and the dependant variable, outcome measurement, can occur simultaneously or the assessment of the outcome is carried out soon after the model induction, (e.g. often the case in forced swim test experiments). The extent to which these experiments have external validity or the extent to which these findings can be generalised to other species, particularly the human condition, can be brought into question (Walker et al., 2014). A clinical diagnosis of depression requires that symptoms be present for at least two weeks (see section 1.2). When an animal model of depression is induced simultaneously with the outcome measure assessment, this brings into question how predictive this model is of an antidepressant effect in the human condition. This concept is discussed further in section 1.6.3 Translational Failure.

One proposal to improve translational failure is by incorporating the entire illness trajectory into the preclinical modelling of depression (McGorry et al., 2006). Currently, few animal models attempt to mimic different stages of the progression of depression, acknowledging that different symptoms may emerge at separate stages of the disease (Walker et al., 2014). This may be one of the reasons why recent compounds have been found to display great promise at the preclinical phase, but have failed to translate successfully to clinical trial (e.g. NK-1 receptor antagonist, Belzung, 2014; Cryan & Slattery, 2007). ‘Back-translation’, or the use of information from clinical studies to improve the preclinical modelling of diseases, was improve translational failure in depression. The use of clinical or epidemiological data on the progression of depression through the course of the human lifetime could improve the translation of preclinical data. One way to get a step closer to stage-specific animal models could be to develop metrics or measures, based on existing outcome measures, such that performance on an outcome measure is indicative of which stage of the disease an animal is displaying (Walker et al., 2014). For example, grading a recognised outcome measure, such as intracranial self-stimulation for anhedonic behaviour, and assigning ‘intensity-dependant’ levels or cut-offs in the outcome measure to particular stages of the disease may be worthwhile research (Walker et al., 2014). The aim of this

approach is to guide phase-specific interventions clinically, which could slow or halt disease progression. Further, developing or targeting specific drugs for acute depression or suicidal patients, interventions to ensure patients experience a longer remission period, or drugs for maintaining remission in between depressive episodes, could result in benefits to the treatment and wellbeing of patients suffering from depression. By modelling depression at a stage-specific level, this could improve the translation of drug treatment findings into clinical trials (Walker et al., 2014). One potential limitation of this approach is the quality of the clinical data available. If these data do not fully characterise the progression of depression or there isn't sufficient data, then this could impact the success of the modelling preclinically (see section 1.3 and 1.6.2).

Another potential method for improving the external validity of depression models by more closely mimicking the human disease is to characterise sub-types of depression within the preclinical data. An example of this would be performing clinical meta-analyses that characterise symptom-clusters to guide preclinical research. Symptom-clusters are defined as "two or more symptoms that are related to each other and that occur together" (Kim et al., 2005, pg. 278). Characterising and successfully modelling several different symptom-clusters preclinically that commonly occur in the human disease could prove a fruitful avenue to test novel compounds to create more targeted or tailored drug treatments. This method could produce higher remission rates in patients when potential treatments are translated through to clinical trials. Further investigation into 'symptom-clusters' clinically could generate knowledge about the different underlying mechanisms leading to symptoms or why some symptoms frequently co-occur (Fried, 2017). See above section 1.3 for further discussion.

As presented above, animal models aim to help us understand the underlying mechanisms behind a disorder, with increased experimental control, and to investigate potential treatments. By assessing the validity of animal models and experiments, we can evaluate whether we are getting as much information as possible from animal models and assess the potential for improvement to model the disorder as best we can in animals.

1.6.2.1 Internal Validity

Internal validity refers to the fact that "*differences observed between groups of animals allocated to different interventions may, apart from random error, be attributed to the treatment under investigation*" (van der Worp et al., 2010, pg. 2). Several biases can threaten the internal validity of an experiment namely; selection bias, performance bias, detection bias, and attrition bias. Selection bias occurs when the allocation of animals to treatment groups is biased. The main method to combat this is to use an unbiased method to randomly allocate animals to groups, preventing any conscious or unconscious bias arising from experimenters manually selecting the animals (van der Worp et al., 2010). One tool to help researchers randomise is the National Centre for the Replacement, Refinement & Reduction of Animals in Research (NC3Rs) Experimental Design Assistant (https://eda.nc3rs.org.uk/) to generate a spreadsheet with the randomised allocation report for experiments. Performance bias occurs when bias is introduced when an experimenter handles the animals differentially during care or under experimental conditions. The main method to combat this is to blind the experimenters to the group the animals belong to, where possible, so that experimenters do not consciously or unconsciously handle animals from separate groups differently (van der Worp et al., 2010). Detection bias or observer bias occurs when bias is introduced into the results when an experimenter is assessing the outcome. This bias may be introduced if the experimenter has knowledge of the group or treatment allocation. One of the main methods to combat this issue is to ensure the experimenter is blinded to the group or treatment allocation during outcome assessment (van der Worp et al., 2010). Attrition bias occurs when drop-outs from experiments or exclusions from the experiments are handled in a biased manner or introduce bias into the results (de Vries et al., 2014; Hoojmans et al., 2014). One of the main methods to combat this is by pre-specifying how and on what basis animals will be excluded from analyses and if possible, or to state *a priori* an intention-to-treat analysis plan where all subjects are included in the final analysis even if there are drop-outs (de Vries et al., 2014; Landis et al., 2012).

It is important to assess the extent to which these measures to reduce the risk of bias are reported in preclinical models of depression to assess the reported quality of these experiments. For example, if reporting of measures to reduce the risk of bias in the literature pertaining to a particular drug is poor, then that provides a rationale for

further studies with higher quality to be conducted prior to translating the findings through to a clinical trial.

1.6.1.2 External Validity

External validity refers to the generalisability of a study to a broader population beyond the animals investigated in the study. There are several areas to note when assessing the external validity of a preclinical study and its generalizability to the human disease namely;

1) differences between animals and humans, differences in comorbidities assessed in the animal literature and the comorbidities likely experienced by patients,
2) differences in the timing of treatment administration and the dosing of treatment between preclinical and clinical situations,
3) differences between the selected outcome measures at the preclinical level and the comparability of these with the outcome measures used in clinical trials.

Traditionally, face validity, predictive validity, and construct validity have been important factors in the development of animal models of depression (Willner, 1984). According to the original criteria of evaluation proposed by Willner (1984), face validity refers to whether the model resembles depression clinically, and whether the antidepressant effects are seen after similar chronic administration, similar to antidepressant efficacy in humans. The concept of face validity can be linked to the diagnostic criteria (ICD-10 & DSM criteria), which rely on cognitive and behavioural dimensions (Belzung & Lemoine, 2011). This brings into consideration a debate of whether face validity should assess how well models mimic a single symptom or whether face validity should refer to how well the model mimics the whole disorder (Belzung & Lemoine, 2011; Willner & Mitchell, 2002). If a model mimics a single symptom, this bears the assumption that the symptoms of a disorder are independent. Currently, research is being carried out to understand the clustering of symptoms in humans and to what extent symptoms are dependent or causally linked (see above section 1.3).

Predictive validity refers to the pharmacological effects of drugs, whether a model of depression is sensitive and specific to drugs that have an antidepressant effect in humans, and whether the dosage required for efficacy reflects the doses given clinically (Willner, 1984). Geyer and Markou (1995) added an important dimension to predictive validity, namely that it should allow one to make *"predictions about the human phenomenon based on the performance of the model"* (Geyer & Markou, 1995, pg 790*)*, which does not limit the definition to the predictability of a drug but expands it also to include underlying psychopathology. As Belzung and Lemoine (2011) note, a definition based only on pharmacological potency relies on "knowing" what effect a drug has in humans.

Broadly speaking, construct validity can be defined as "the accuracy with which the model measures what it is intended to measure" (Koob, Heinrichs & Britton, 1998). Further it can be defined as the theoretical relationship to depression (Willner, 1984). As Willner (1994) expanded on in a subsequent publication, this includes the similarity with the human etiology, biological dysfunctions, the cyclical nature of the disorder, and information pertaining to events that trigger depression, as well as the link between the etiology and the dysfunctions (Belzung & Lemoine, 2011).

In addition to assessing these main types of validity, Belzung and colleagues have worked extensively to further characterise aspects of the animal models of neuropsychiatry that can assist to improve animal models of neuropsychiatry (Belzung & Lemoine, 2011; Willner, 1994). Their proposed new framework broadly encompasses the same three main criteria; construct validity, predictive validity and face validity, in much more detail with sub-categories of validity within these broad groups, as well as incorporating new concepts such as mechanistic validity, biomarker validity, and induction validity. These newly defined concepts will be discussed in more detail below and are highlighted in the figure from Belzung and Lemoine (2011).

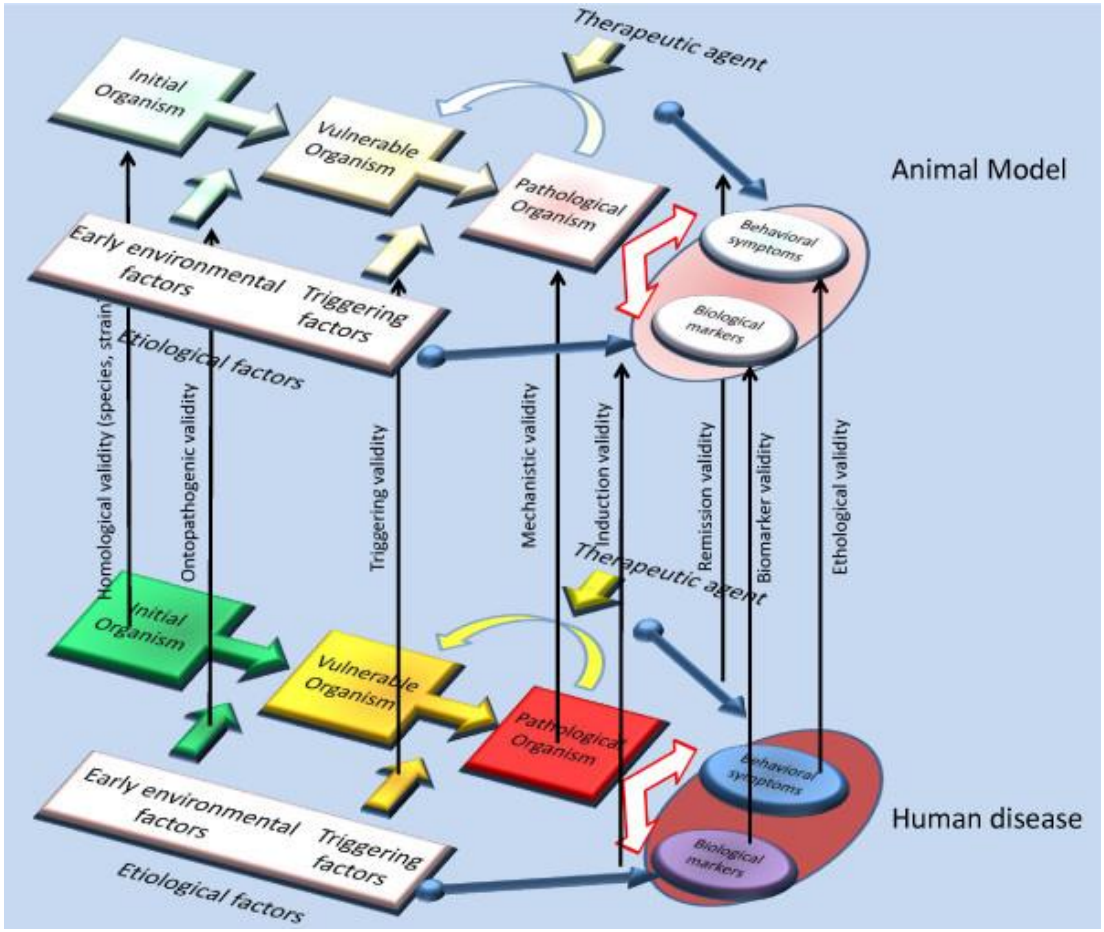


Figure 1.3. "Criteria of validity for animal models" (Belzung & Lemoine, 2011, pg. 8)

Firstly, homological validity assesses which strain and species have been chosen for the model. This concept takes into consideration the biological and anatomical similarity between the species used as a model and human biological mechanisms (Belzung & Lemoine, 2011). Pathological validity assesses "*the similarity of processes that lead to disease*" (Belzung & Lemoine, 2011, pg. 7). This concept encompasses two sub-categories of validity, ontopathogenic validity, which refers to whether similar environmental factors in early life produce a vulnerable state in the organism (Belzung & Lemoine, 2011). The second; triggering validity, refers to whether similar factors during adulthood trigger the organism to enter a vulnerable or disease state (Belzung & Lemoine, 2011). Mechanistic validity assesses the "*similarity of the mechanisms we suppose or know in the animal disease, compared to the mechanism that is proposed or known in humans*" (Belzung & Lemoine, 2011, pg. 8). This concept encompasses both the mechanisms that we postulate cause a symptom, as well as the mechanisms

through which therapeutic interventions are thought to work. Belzung and Lemoine differentiate between the mechanisms and the effects of the mechanisms. Symptoms may be the result of interactions of several mechanisms, rather than the direct effect of a mechanism (Belzung & Lemoine, 2011).

Secondly, face validity corresponds similarly with Willner's original concept, the referring to the similarity between what is observed in animals and humans, with the addition of explicit sub-concepts; ethological validity and biomarker validity. Ethological validity refers to whether pathological behaviours are similar in the animal model compared to the human disease. Belzung and Lemoine highlight that this needn't be a direct comparison, rather behaviours characteristic of that species (Belzung & Lemoine, 2011). Biomarker validity refers to whether the function of biomarkers is similar across species.

Thirdly, predictive validity refers to "*the similarity of the relationship between the triggering factors and the occurrence of the disease as well as the relationship between the therapeutic agent and the disease state*" (Belzung & Lemoine, 2011, pg. 9). This concept assessed symptoms with broad brush strokes and refers to whether there is a similarity between the impact of etiological factors as well as the observable effects of the treatment investigated without looking at the mechanism that is at work (Belzung & Lemoine, 2011), and encompasses the sub-concepts induction validity and remission validity.

> *" An animal model of disease is not just a model of the action of a therapeutic agent at time point T. It has to draw from the comparison between two pathological organisms, but possibly also mimic the temporal and etiological process of transformation from a healthy organism to a pathological one via the state of vulnerability."*
>
> Belzung & Lemoine (2011, pg 7)

Induction validity refers to whether the etiological factors have similar actions of the observable factors in the outcome measures as in the human condition. Remission validity refers to the same concept, but whether the observable effects of a treatment in an animal model are similar to that seen in the human condition.

There is merit to the added understanding of the strengths and limitations of animal models, as assessed by these criteria. There is a potential lack of understanding of the interface between the psychopathology, the underlying psychological processes behind depression, and the pathophysiological, the biological implications. (Belzung, Willner, Philippot, 2015). How much do we know about the human biology, psychology, and the interaction of these mechanisms, and how far can we map these mechanisms in animal models? How can we best use the data and knowledge from animal models to inform our treatments in the human condition? And how can we use knowledge about the clinical condition to help improve the animal models we currently have?

1.6.3 Broad Categories of Model Induction in Animal Models of Depression

Building on the framework laid out by Belzung and Lemoine (2011), an organism may have a predisposing genetic vulnerability, early environmental factors or triggering factors. These may take the organism from a vulnerable state into a pathological organism, where the cognitive and biological mechanisms are different from the initial organism state.

Animal models of depression can be categorised into models that aim to mimic genetic factors, early environmental factors, and triggering factors. Therefore, these models can be categorised into Genetic Models, Developmental (in early life or adolescence) Models, and Acute Models, respectively. A number of techniques are used to induce depression across these stages. Stress, pharmacological interventions, and genetic manipulations can be implemented across all life stages. Stress models grew out of the Diathesis-Stress hypothesis, which hypothesises that disorders develop out of an interaction between a predisposition to a vulnerable state and a stressful event caused by life experiences (Monroe & Simmons, 1991; Abramson, Metalsky, & Alloy, 1989). Predisposition has expanded from its original

definition of biological or genetic factors, to include psychological and situational factors.

These models exhibit a number of cognitive, behavioural, and biological phenotypes. Genetic Models include transgenic models, and selectively bred models such as Flinders Sensitive Line (FSL) or congenital learned helplessness. Transgenic models are where a specific gene has been manipulated to increase or decrease expression and induce behavioural phenotypes such as learned helplessness behaviour (Wang et al., 2017; Ridder et al., 2005). Most transgenic models alter genes involved in the serotonin and glucocorticoid pathways and help elucidate the mechanisms behind the involvement of these genes and proteins in depression. Selectively bred models take animals that display a certain behavioural phenotype and breed them to maintain the behaviour across generations. Selectively bred animals such as the FSL animals exhibit behavioural despair phenotypes as well as biological phenotypes similar to depressed patients such as cholinergic sensitivity (e.g. Overstreet et al., 1996). Developmental Models include early life stress models such as maternal separation, and the use of pharmacological interventions such as cytokine and corticosterone injections to target the HPA-axis or inflammatory pathways and induce a biological state similar to depression. Gluco-corticosterone injections induce a physiological state of stress with behavioural despair and anhedonia (Zhao et al., 2008). Stress or pharmacological interventions can be administered acutely or chronically to induce a depressive-like phenotype, and they can be administered both in early life, adolescence and adulthood.

Acute models in adulthood aim to mimic the triggering factors that mean an organism transfers into a disease state, either from an initial “healthy” state, or from a vulnerable state. Several different interventions have been developed and implemented in relation to depression. Firstly, models that utilise stress have been implemented including; unpredictable chronic mild stress (Katz et al., 1981) where animals are subjected to a series of different unpredictable stressors over a period of weeks. Further models include chronic restraint stress (Holsboer et al., 1987), where animals are restrained from moving for a number of days or weeks, and social defeat stress where animals are subjected to an emotional stressor of an older, more aggressive animal in the home cage (Blanchard & Blanchard, 1977). Animals exposed to stressful

events display key behavioural phenotypes such as anhedonia and learned helplessness as well as key biological readouts such as changes in body weight, sexual behaviour, and changes in biological readouts from the HPA axis network and neurotransmitter levels (Wang et al., 2017). Additional models include olfactory bulbectomy, a surgical model where removal of the olfactory bulb in rodents elicits biological and behavioural mechanisms that mimic depression (van Reisen, Schnieden & Wren, 1976). Other acute pharmacological models include administering corticosterone, cytokine, or glucocorticoid insults (Wang et al., 2017) either a single administration or chronically over a number of days to stimulate the HPA axis or inflammatory pathways, see above section 1.3 Pathology.

Another commonly administered acute stressor is the forced swim test (FST) in rats or the tail suspension test (TST) in mice. The FST was originally developed by Porsolt and colleagues in 1977 for mice and the TST in 1978 for rats were introduced as models of depression. The FST involves subjecting animals to an acute stressor of being in a water tank from which they are unable to escape. The TST involves hanging mice by the tail above the ground, similar to the FST, a stressful event from which animals are unable to escape (Steru et al., 1985). At first, animals try to escape from the situation, but after some time give up and display behavioural despair. In some experiments, a pre-swim is administered approximately 24 hours prior to the testing, where animals are exposed to the stressful situation for a longer period.

Although sensitive to pharmacological agents, recent critique has characterised this test as an antidepressant screening tool rather than a model of depression in line with the validity criteria set out by Janssen in the 1960s; that the aim of behavioural tests such as these was to find a "*reproducible, reliable and rapid method to test compounds"* (Janssen, 1964). Further critique towards the FST and TST has called into question the cognitive mechanisms behind the behavioural despair, with some authors claiming that despair behaviour reflects coping mechanisms and adaptive behaviour (Commons et al., 2017; Molendijk & de Kloet, 2015). Despite these criticisms, the FST and TST are still widely used to assess behavioural despair in rodents and assess antidepressant efficacy.

1.6.4 Translational Failure

The aim of using animal models is to help understand the human disorder and to extensively test therapeutic agents prior to application in humans. To maximise information gained from animal models in depression, it is paramount to have a clear framework with which to understand our work. The field of neuropsychiatry has many methods of inducing depressive-like phenotypes, all of which contribute unique knowledge, however, attempts to systematically synthesise this evidence in light of a clear framework has been lacking, as well as concerns regarding the quality of preclinical animal studies.

Even with seemingly good face validity and predictive validity, preclinical findings, do not always translate successfully into clinical trials. This has been seen especially in the field of drug interventions for depression (Cryan & Slattery, 2007). One example of poor translation is the case of NK-1 antagonists for depression. NK-1 antagonists displayed very promising results in reducing depressive-like behaviour in preclinical trials and were tested thoroughly, with various models and in a range of outcome measures (Belzung, 2014). This potential treatment was taken through clinical trial with mixed results in Phase 2 and ultimately discontinued. The following section will go into more detail about the potential reasons for translational failure.

There are some potential pitfalls when translating promising findings from preclinical studies into clinical trials. These fall generally into three categories;

1) limited internal validity of the preclinical studies,

2) limited external validity of the preclinical studies, and

3) poor design of clinical trials (van der Worp et al., 2011).

Firstly, internal validity, as highlighted above, has important ramifications for the confidence in the findings from a study. The investigation into the reporting of measures to reduce the risk of bias in other neurological fields has revealed that studies that do not report measures to reduce the risk of bias tend to overstate the efficacy of a drug (Van der Worp et al., 2007; Van der Worp et al., 2010). This might

translate to animal models of depression both for the overstatement of efficacy of treatments as well as an overstatement of the evidence for biological mechanisms altered with knock-out animals for example.

Secondly, external validity, the confidence with which we can generalise the findings of a study to a broader population beyond the subjects investigated in the study. Some original criteria by which to evaluate the external validity of animal models of depression can be critiqued. The idea of predictive validity that is based on pharmacological efficacy similar to human efficacy, can be seen as a problematic argument, in that the efficacy must be known in a human population. For the development of new drugs and alteration of mechanisms thought to be involved in depressive-phenotypes, the mechanisms must be known in the human population. Further, the use of screening assays such as the FST and the TST can be critiqued. When interventions used to induce depressive-like phenotypes are simultaneously used as outcome measure assessment, this brings into question how predictive this model is of an antidepressant effect in the human condition. Further, the acute administration of antidepressant agents reverses the behavioural despair in FST and TST, whereas these drugs take weeks to be effective in humans (Krishnan & Nestler, 2011). These are just examples of potential issues that could impact the translation of findings from animals to inform human treatments and research.

Thirdly, the final issue when translating preclinical findings to clinical trial is the potential for key limitations of preclinical literature to be overlooked when designing the clinical trial (van der Worp et al., 2010). Potential ways to tackle this would be to have inter-disciplinary groups involved when translating preclinical research, or for thorough systematic reviews of the preclinical literature to be conducted prior to the clinical trial design.

Therefore, it is of interest to systematically gather information about animal models of depression, both to provide an overview of the field as well as assess the reported quality of these studies. With regard to external validity, which animal models do we have most confidence in? What is the quality of evidence available from different animal models? What types of models that are used in further understanding the

mechanisms behind depression by modelling phenotypes in animals and experiments used to understand antidepressant drug development? Is there evidence that these differences could this affect the success of subsequent translation?

The aims and objectives of this thesis are outlined below.

1.7 Aims & Objectives

Overall Aim: To provide an overview of the *in vivo* animal models of depression used in the literature.

Objectives:

1. **Model induction:** Investigate how animal models of depression are induced. Which models provide the most reliable outcome yet cause least suffering to animals?
2. **Outcome measures:**
    a. Commonly used outcome measures: Investigate the outcome measures assessed in animal models of depression. What are the relationships between outcome measures? Are there tests which are more reliable to measure an outcome? Investigate the cost-effectiveness of outcome measures, are there any commonly used tests which can be streamlined (e.g. in severity or number of animals used) and still produce a reliable/significant effect?
    b. Outcome measures of clinical relevance: Are the outcomes measured in animal models significantly relevant to the endpoints investigated in human trials? Can translatability be increased? Do behaviours induced in animals reflect the data-driven clusters of symptoms in patients?
3. **Quality of studies:** Investigate reporting bias and quality in the in vivo modelling of depression literature. What is the standard of reporting and what measures are commonly implemented in the preclinical depression literature

to reduce bias? How can bias in the preclinical depression literature be reduced?

4. **Drug efficacy:** Provide an overview of the efficacy of different drug interventions on symptom-reduction in *in vivo* animal models.

2 SYSTEMATIC REVIEW & META-ANALYSIS

2.1 Systematic Review

"Large-scale review and integration of existing theory and research may be considered a type of research in its own right - one using a characteristic set of research techniques." - Feldman, 1971, pg. 86

Systematic review is a technique used to systematically search for and gather published literature to answer a specific research question. This technique was coined by Feldman in 1971 when he wrote that this technique should be a research field of its own. The use of this technique has increased since the term was coined, as is reflected in the amount of papers published with this key term. A PubMed search yielded 5719 results for 'systematic review' in 1971 and over 120,000 hits in 2017. The Cochrane Collaboration has been fundamental in the setting up of gold standards; a framework for systematic review methodology and advocating for the use of these techniques to inform healthcare policy (Vesterinen et al., 2014). Whilst Cochrane have helped to revolutionise the use of systematic review and its wide-spread use, their techniques are primarily for the analysis of interventions in humans and there are considerations that need to be taken into account when transferring the use of these tools for systematically reviewing and analysing data from animal studies.

Systematic review (SR) is different to other literature synthesis techniques in that the formal steps of the process are explicitly defined so they can be reproduced by other teams (Biondi-Zoccai et al., 2011). The formal steps involve; extensive searching for primary studies on a topic, using explicit criteria for selecting studies, and criteria for appraising the internal validity of studies, indentifying what data or information will be extracted from the studies, and where appropriate the pooling of statistical data, and finally, publishing the findings. The steps taken for a project are all outlined in a prespecified protocol to ensure the process is transparent and clearly defined. These will be discussed below.

Figure 2.1. Steps of a Systematic Review

2.1.1 Research Question

Forming the research question is the first step of SR and outlines the rest of the process. Research questions can ask a specific, usually healthcare related, question or simply provide an overview of a field. In SRs of animal studies, research questions may relate to a) understanding which variables impact the effectiveness of an intervention to inform the design of future animal studies; as well as b) getting an overview of the type, breadth and quality of evidence about an intervention to assess whether there is enough evidence to translate the findings to clinical trials.

2.1.2 Search Strategy

The aim is to identify all available literature pertaining to your research question. Systematically searching for literature relevant to the research question often involves creating search strings to interrogate online bibliographic repositories or databases of published studies. Repositories such as PubMed, EMBASE, Web of Science, and PSYCHinfo are commonly searched. Systematic searching may also involve searching through grey literature and unpublished literature such as in-house animal trials done in pharmaceutical companies, or hand searching the reference lists of studies included in the review. The methods for identifying literature should be explicitly described. Systematic Review Centre for Laboratory Animal Experimentation (SYRCLE) has developed extensive search strings for PubMed and EMBASE that encompass animals and animal experimentation (Leenaars et al., 2012). These search strings can be added to the disease/drug search string to try and identify those articles that report experiments in animals.

2.1.3 Study Selection

Criteria for selecting studies to be included in the systematic review are very specific to the field and to the research question. Inclusion criteria often include key terms included in the research question, such as whether the paper mentions the use of the drug of interest, the disease model of interest, and the experiment performed on the species of interest. Inclusion or exclusion criteria may also relate to the experimental design used in the study, the dates the study was published, the language in which the study was published, and whether the paper mentions key pieces of information such as sample sizes or measures to reduce the risk of bias. These criteria must be clearly defined and pre-specified to reduce bias in the review where studies are included based on significant findings and reduce the subjectivity of the independent reviewers (de Vries et al., 2015). It may be necesary to apply different inclusion criteria at title and abstract screening stage and at the full-text screening stage because there may not be sufficient information in the abstract to make a decision as to whether an article is relevant or not.

2.1.4 Criteria for appraising internal validity of studies

It is relevant to assess the quality of the primary studies that are included in the systematic review, to know the strength of evidence. The criteria with which internal validity is assessed, discussed more in depth in the previous chapter, involves the reporting of measures to reduce the risk of bias such as whether subjects were randomly allocated to groups, whether the experimenter was blinded to the group allocation for the duration of the experiment and during outcome assessment. Criteria upon which to assess measures to reduce the risk of bias may include, the CAMARADES checklist (Macleod et al., 2004), the Landis 4 criteria (Landis et al., 2012), SYRCLE risk of bias tool (Hooijmans et al., 2014), and aspects of the ARRIVE guidelines (Kilkenny et al., 2010).

2.1.5 Data Extraction

In the protocol of the SR, the information and data that will be extracted from all primary articles, where reported, are specified. Information extracted may relate to factors of experimental design thought to be relevant to the impact of the intervention

or drug. If a quantitative summary of the data is planned, key values such as group sample size, the key statistic of interest, and a measure of variance for each group may be extracted to compute a summary effect size. All data extraction is done in duplicate; therefore it is key to pre-define what aspects will be extracted.

2.1.6 Meta-Analysis

Where appropriate, it may be possible to pool the statistical data from the primary articles to get an overall estimate of whether an intervention is effective and explore the variables that impact the effectiveness of the intervention. It might further be of interest to assess the degree of publication bias within a field. These analyses are conducted using meta-analysis (MA) and will be discussed in more depth below.

2.1.7 Publish

The final stage of the SR process is to publish the findings and disseminate the knowledge to ensure that the key stakeholders and consumers of the research, whether policymakers, primary animal researchers, or clinical trialists, have access. The report of the SR and MA should include all the necessary information for readers to assess how the systematic review has been carried out and assess the measures to reduce the risk of bias in the SR in compliance with the PRISMA guidelines (Liberati et al., 2009). Just as it is important to assess the quality of primary studies included in a systematic review, it is just as important to ensure that systematic reviews are reported adequately and that methods are transparent and reproducible. Several tools for assessing the reporting of systematic reviews, such as the Oxman and Guyate Index for appraising systematic reviews (Oxman & Guyatt, 1991) which spurred the development of the AMSTAR tool (Shea et al., 2009) to assess the quality of systematic reviews, and guidelines for authors of SRs to ensure transparent reporting, the PRISMA guidelines (Liberati et al., 2009). A similar guide has been developed for authors of systematic reviews of animal models (Sena et al., 2014).

2.2 Meta-Analysis

"I like to think of the meta-analytic process as similar to being in a helicopter. On the ground individual trees are visible with high resolution. This resolution diminishes as the helicopter rises, and in its place we begin to see patterns not visible from the ground"

- Ingram Olkin, quoted in Biondi-Zoccai et al., 2011

Meta-analysis is a statistical technique used to quantitatively summarise the primary data, often by summarising the effects for each study or experiment. The estimation of effect sizes from several studies is a method that has been used since Pearson in 1904. Pearson estimated the average correlation between inoculation for enteric fever and mortality. This was applied to other fields in 1931 when Tippet compared different farming techniques on agricultural yield. The term 'meta-analysis' was first coined by Glass and Smith in 1976 when they introduced statistical techniques for pooling data from several studies that had been collected systematically. The statistical tools of meta-analysis are dependent on the type of primary data that you are trying to pool. As the data from and characteristics of animal studies differ from clinical studies, different effect size calculations are used, and different considerations need to be taken.

The first step of meta-analysis involves extracting the statistics from the primary articles of interest to calculate an effect size for each study, the mean or correlation, the number of subjects per group and an estimate of variance from each group. Several types of effect sizes can be calculated based on the data in the primary studies, including the effect sizes of means, correlational data, and effect sizes based on binary data such as risk ratio and odds ratio. The effect sizes from each study are pooled together under one of two assumptions, based on whether the true effect size is assumed to be the same in all studies or whether the true effect sizes differ between studies (Borenstein et al., 2009). The pooled effect size allows us to understand the direction of the effect and the magnitude of the effect (de Vries et al., 2015). Further, we can investigate the heterogeneity, the variation in true effects between studies, to "make sense of the pattern of effects" (Borenstein et al., 2009, pg. 107). Firstly, heterogeneity can be quantified using various statistics (see Borenstein et al., 2009

and Vesterinen et al., 2014). Then differences between studies are explored with, for example, sub-group analyses or meta-regression to assess the relationship between study-level variables and the effect size (Borenstein et al., 2009).

Borenstein and colleagues provide an extensive guide to meta-analysis in their book published in 2009. Further they provide workshop and training materials to researchers embarking on this research endeavour (https://www.meta-analysis.com/). For a more specific guide on meta-analysis of data from animal studies, see the article by Vesterinen and colleagues (2014), which clearly outlines the considerations necessary for applying this technique to data from animal studies. Specific equations used in the below met-analyses will be explicitly stated in the relevant results chapters (Chapter 5 microbiota systematic, Chapter 6 Ketamine systematic review).

2.3 Systematic Review and Meta-Analysis of Animal Studies

The field of applying systematic review and meta-analysis to animal modelling literature first started pooling data from experimental stroke studies in the early 2000s. In 2001 Horn and colleagues investigated nimodipine in experimental stroke, and later in 2004 Macleod and colleagues investigated the effects of nicotinamide on experimental stroke (Macleod et al., 2004) and O'Collins and colleagues investigated evidence across interventions that target stroke in animal models (2006). The Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies (CAMARADES) was founded that year in 2004. A research group with similar efforts is the Systematic Review Centre for Laboratory Animal Experimentation (SYRCLE), which was established in 2008. Since the initial application of these tools to the field of stroke in 2001 and 2004, they have been applied to many other neurological fields including Alzheimer's disease (Egan et al., 2016), breast cancer (Chen et al., 2016), glioma (Jue et al., 2018), spinal cord injury (Watzlawick et al., 2016), Parkinson's disease (Rooke et al., 2011), multiple sclerosis (Vesterinen et al., 2010), pain (Currie et al., 2013; Seretny et al., 2014), and schizophrenia (Bahor et al., 2016). It has not yet been applied to the field of animal models of depression.

SRs of animal studies are exploratory and hypothesis-generating (Vesterinen et al., 2014). All available relevant studies are included in a review and the impact of experimental design variables and sources of bias are explored to inform the design of future animal studies and of clinical trials (Vesterinen et al., 2014). Due to the sheer volume of preclinical investigations of depression continually being published, it is difficult to achieve an overview of what is already known, and to assess the marginal contribution of new research (de Vries et al., 2011). In this context, a systematic review of the existing preclinical literature can provide an unbiased, collective overview of existing knowledge which can help avoid unnecessary replication of good quality evidence as well as highlight gaps in knowledge and areas for additional, higher quality evidence to inform future experiments (Vesterinen et al., 2014). It can also provide a broader overview and understanding of the laboratory methods used to induce depression, the range of outcome measures used to assess depressive-like phenotypes, and the variables that might impact on the efficacy of different treatments (de Vries et al., 2011). Providing a broad overview of the literature can help avoid unnecessary replication of good quality animal studies (Vesterinen et al., 2014). This information is valuable to translation of findings through to clinical trials and can inform trial design (Vesterinen et al., 2014). The findings from preclinical systematic review and meta-analysis may also contribute to the refinement of methods used in animal investigations of depression, reducing the distress caused to animals by substitution with equally informative methods of lower severity; contribute to the optimisation of the numbers of animals used in depression research by informing well founded power calculations.

The nature and design of primary animal studies means that the conduct and interpretation of meta-analysis is different to the established meta-analysis of clinical trial data. Animal studies tend to have small sample sizes, with studies being conducted across many laboratories introducing variance (Vesterinen et al., 2014). Even with the use of transgenic animals, the impact of sex of animals, animal husbandry and other experimental design variables mean there are more differences between studies in SRs of animal data vs SRs of clinical trials. Therefore, investigating heterogeneity, the differences between studies, is a primary aim of meta-analysis of animal data. Further, there is more lack of independence both within studies and between studies. In a single study there may be several outcomes from the same

groups of animals, as well as a single control group serving multiple treatment groups, which contribute to the lack of independence within studies. Several studies may have been conducted at the same laboratory or by the same experimenter, which contributes to the lack of independence between studies. Nesting of outcomes from the same animals is the primary technique used to combat the lack of within study independence (Vesterinen et al., 2014). Hedges and colleagues have proposed a number of further statistical tools to deal with the lack of independence between studies (Hedges et al., 2010).

2.3.1 Limitations of Systematic Review and Meta-Analysis Methods

Despite its merits, there are a number of limitations of this methodology. Firstly, statistical limitations of meta-analysis, including the quality of the studies included. Secondly, bias in the sampling and issues of publication bias. Thirdly, the timeliness of the findings from the review. These issues, discussed below, can influence the overall results and reliability of a systematic review and meta-analysis.

The findings from MA can be critiqued in that no meaningful information can be gained from averaging data from studies that have different definitions of the dependent or independent variable, or similar research questions (Hunt, 1997 in Rosenthal & Dimatteo, 2001). However, Rosenthal and DiMatteo argue that a quantitative summary can be useful when generalising findings to a field. In disease modelling in animals, it can be useful to understand the effects of an intervention in different circumstances, for example in different species, to evaluate the generalisability of findings.

Meta-analysis is dependent on the quality of studies. In SRs of randomised controlled trials, poor primary study quality may be an exclusion criterion. In SRs of animal studies, often we don't have this option as the reporting of measures to reduce the risk of bias varies greatly, and many primary studies are of poor quality (Sena et al., 2007; Rosenthal & DiMatteo, 2001). There are differences in the reported effectiveness of treatments, were studies that do report measures such as, whether animals have been randomly assigned to treatment or control group and whether the

experimenter was blinded to the group assignment during outcome assessment, tend to report a more conservative effect size measure compared to those that do not report these measures (Macleod et al., 2015). It is important to assess the reporting of measures to reduce the risk of bias, as SR is also a form of observational research, we cannot exclude these factors from having an impact on the reported effectiveness of an intervention and we therefore we can only assess what has been reported (Vesterinen et al., 2014).

Issues of non-independence of effects between and within studies, as discussed above, are dealt with using nesting techniques and other statistical tools (Vesterinen et al., 2014; Hedges et al., 2010). We must be sure to assess systematic reviews and meta-analyses with stringent criteria to ensure the quality of the findings, including ensuring that meta-analyses are adequately powered to detect pooled effects (Wang et al., 2018). Adequate power ensures that the findings from a study are not likely to be falsely positive, and therefore we conclude wrongly from the analysis. In meta-analysis, statistical tests are used to determine whether there is significant heterogeneity between studies. If a meta-analysis is underpowered, there is a higher chance that a significant result in heterogeneity tests is falsely positive (Wang et al., 2018). Simulation studies by Wang and colleagues found that meta-analyses using normalised mean difference to calculate effect size were higher powered than when calculating effect size using standardised mean difference (Wang et al., 2018). Further, they assessed the power of methods to explore heterogeneity. They found that when exploring heterogeneity, meta-regression was preferred with lower false positive rates compared to stratified meta-analysis (Wang et al., 2018). Quantitative methods for assessing publication bias can also be assessed for appropriateness. Funnel plots are the technique used for plotting of effect sizes against precision, where identification of asymmetry can indicate small study effects and publication bias. Zwetsloot and colleagues found that using standardised mean difference as your measure of effect size in funnel plots can distort the funnel plot and cause over-estimation of the presence of publication bias (Zwetsloot et al., 2017). Authors recommend for meta-analyses where the use of standardised mean difference for calculating effect size is required, to use sample size based precision estimates. Using the most appropriate statistical methodology for meta-analysis, and

acknowledging known pitfalls, ensures that findings from SR and MA are as robust as possible.

Systematic reviews can only present findings that are accessible. There will inevitably be studies that are missed during the search, however publication bias is rife across the field of biomedical sciences (Sena et al., 2010; Rosenthal & DiMatteo, 2001). Publication bias occurs when published studies are more likely to find their way into a meta-analysis, combined with the fact that significant studies are more likely to be published (Borenstein et al., 2009). If the studies that are not published have different findings or conclusions to those that are published, this introduces bias both in systematic reviews of the literature and in general searches of the online databases, and therefore the current understanding of a field can be biased (Sena et al., 2010).

One approach that can potentially improve publication bias as well as addressing the internal validity of studies is the Open Science movement. The Open Science movement has been trying to address some of these issues we see in animal studies, by advocating to make scientific research and data accessible to all. They campaign actively for open access publications to make research more easily identifiable, available and make the reporting of research more transparent. They facilitate the use of open tools to make this process easier, including open notebook science and open software (UNESCO, 2017).

"Sharing data and materials signals that researchers value transparency and have confidence in their own research." - (McKiernan et al., 2016)

Open access research has huge benefits to the community in that patients can access research to understand benefits and limitations of available treatments, tax-payers can access research that has been paid for with their contributions. Further benefits for researchers are that open access research is used more and has higher impact including, increased citation (Harnad, 2007; McKiernan et al., 2016; Wagner, 2010). Open access research is twice as likely to be published (Harnad, 2007). Not least, a major benefit of open access is the increased access for other researchers, including

meta-researchers, allowing more research to be available and included in evidence synthesis to provide the fullest picture of the evidence (Harmad, 2007).

The practise of uploading partial or completed studies to preprint servers and experimental data from these studies to repositories has also increased (Lin, 2018). This practise has a number of benefits; encouraging the linking of data and pre-specified analyses can improve the internal validity of studies, the content is accessible in future both for researchers and for data requests (McKiernan et al., 2016). This overall increased transparency and increase in access of data can ensure that all available data is included in systematic summaries of a research topic providing a more accurate estimate in meta-analysis.

Olkin noted over 25 years ago that a repository where data from published papers would be ideal to solve many of these issues, but that in 1995 was unattainable at the current date (1995). The field has made progress since Olkin's remarks. A recent culture shift from the International Committee of Medical Journal Editors (ICMJE) comes in the form of a proposal to require data generated by interventional clinical trials that are published in its member journals to be responsibly shared with external investigators. Many large funders are supportive, with support from industry and academic trialists as well (Ross, 2016).

Further, the registration of preclinical trials has been proposed as another method that would increase the internal validity of animal studies and ensure that published findings are in keeping with initial primary hypotheses, that exploratory research is not published as hypothesis-testing confirmatory experiments. A set-up similar to the proposal by the ICMJE, requiring hypothesis-testing animal experiments that investigate an intervention to responsibly share the data, at least for the primary hypothesis, would also aid in increasing transparency and aim to increase the quality of published experiments (Naci et al., 2015).

Not all stakeholders share this viewpoint and enthusiasm for the topic. Researchers are apprehensive that competitive data that may form the basis for patents or

intellectual property rights may cause research groups to miss out or being “scooped” (Flather, 2015). Further concerns include data being analysed by teams or individuals without statistical expertise, and that post-hoc analyses are performed without sufficient statistical power. Further, although open access is growing in popularity, the widespread adoption of these practices has not yet caught on (McKiernan et al., 2016).

The final limitation of systematic review and meta-analysis is that they are resource consuming research activities. The larger the review, the more studies are involved in the process, the longer the process takes. An average Cochrane review has been estimated to take approximately 67 weeks (Borah et al., 2017). This means that findings from large systematic reviews are likely to be out of date when published. The increasing amounts of literature published in biomedical sciences make the process of systematic review and meta-analysis time consuming. Olkin remarked in 1995, that meta-analysis is the key to dealing with the increasing amounts of information. However, this era of information explosion has further erupted, and novel tools and approaches are required to help sort and categorise all of the information. One technique to assist with this limitation is the application of automation tools. Machine-learning and text-mining tools can be applied to steps of the systematic review to help reduce the amount of human resources required. These will be discussed further in Chapter 3 and Chapter 4.

The methodology for this broad systematic review and meta-analysis is outlined below with a pre-specified protocol published in 2016.

2.4 Protocol for the Systematic Review

Received: 20 July 2016 | Revised: 24 November 2016 | Accepted: 10 January 2017
DOI: 10.1002/ebm2.24

WILEY Evidence-based Preclinical Medicine

SYSTEMATIC REVIEW PROTOCOL

Understanding *in vivo* modelling of depression in non-human animals: a systematic review protocol

Alexandra Bannach-Brown[1,2] | Jing Liao[2] | Gregers Wegener[1] | Malcolm Macleod[2]

[1]Aarhus Universitet, Aarhus C, Denmark

[2]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh UK

Correspondence
A. Bannach-Brown, Clinical Neurosciences, University of Edinburgh, Edinburgh, UK.
Email: a.bannach-brown@ed.ac.uk

Funding information
National Centre for the Replacement, Refinement & Reduction of Animals in Research, Grant/Award number: (NC/L000970/1).

The aim of this study is to systematically collect all published preclinical non-human animal literature on depression to provide an unbiased overview of existing knowledge. A systematic search will be carried out in PubMed and Embase. Studies will be included if they use non-human animal experimental model(s) to induce or mimic a depressive-like phenotype. Data that will be extracted include the model or method of induction; species and gender of the animals used; the behavioural, anatomical, electrophysiological, neurochemical or genetic outcome measure(s) used; risk of bias/quality of reporting; and any intervention(s) tested. There were no exclusion criteria based on language or date of publication. Automation techniques will be used, where appropriate, to reduce the human reviewer time. Meta-analyses will be conducted if feasible. This broad systematic review aims to gain a better understanding of the strengths and limitations of current approaches, models and outcome measures used. This study aims to provide insights into factors affecting the efficiency of model induction and the efficacy of intervention. Here, we outline the protocol for a systematic review and possible meta-analysis of the preclinical studies modelling depression-like behaviours and phenotypes in animals.

KEYWORDS
animal models, depression, risk of bias, validity

1 | BACKGROUND

1.1 | What is already known about this disease/model/intervention? Why is it important to do this review?

Depression is a mental illness characterized by *"low mood, loss of interest and pleasure or loss of energy."*[1] It is the leading cause of disability in the world[2] and is currently the brain disorder with the highest financial cost in Europe.[3] The number of people diagnosed with depression worldwide is estimated to be 400 million.[4] Depression places a huge burden on patients and poses a great cost to healthcare systems and governments. The rate of remission with antidepressant medication is, at best, 70% and may only be achieved after several levels of intervention.[5,6] Despite decades of investigation into depression, little is known about the biological mechanisms underpinning the disease.[7,8] With better understanding of the mechanisms causing depression, the development of novel and more reliable treatments might be possible. There is solid rationale that further investigation into the mechanisms and factors that contribute to the development of depression is needed. This is a highly important area to tackle, both from a clinical and a preclinical perspective.

Preclinical investigations contribute significantly to understanding the mechanisms underlying depression, which can, in turn, inform treatment development and increase translational success of clinical research. One example of this contribution is the systematic review and meta-analysis of antidepressants for the treatment of stroke.[9] Analysis of the evidence in the preclinical animal literature informed key aspects of the trial design in the subsequent FOCUS trial.[10] Preclinical experiments have the ability to model and dissect important mechanisms of depression and therefore provide insights into the neurobiology of the disorder.[11] Preclinical experiments can also investigate the safety and efficacy of proposed treatments prior to exposure in human cohorts.[12] The knowledge from preclinical

investigations can aid prevention research, translating findings into the best and earliest interventions for the human disease, which are top research priorities recently identified by an MQ: Transforming Mental Health report.[13]

Due to the sheer volume of preclinical investigations of depression, it is difficult to achieve an overview of what is already known and to assess the marginal contribution of new research.[23] In this context, a systematic review of the existing preclinical literature could provide an unbiased, collective overview of existing knowledge and allow the additional contribution of new research to be assessed. It could also provide better understanding of the laboratory methods used to induce the condition, the range of outcome measures used to assess depression phenotypes and the variables that might impact the efficacy of different treatments.[14] The findings from this systematic review and meta-analysis may also contribute to the refinement of methods used in animal investigations of depression, reducing the distress caused to animals by substitution with equally informative methods of lower severity, and contribute to the optimisation of the numbers of animals used in depression research by informing well-founded power calculations.

What is meant by an experimental non-human animal model of depression? An experimental model in preclinical non-human animal neuropsychiatric research is defined as including both a dependant variable (an outcome measurement) as well as an independent variable (model induction or manipulation).[15] We differentiate between 3 broad experimental designs within the animal depression literature and anti-depressant drug literature:

1. Studies which compare a control group to a group of animals that receive a lesion (model induction) on an outcome measure. These studies may also have a drug arm.
2. Studies which only compare a "lesion" group of animals to a group of lesioned animals that receive a drug intervention. Once a lesion method has been sufficiently established (known to be valid and reliable), experimenters know a lesion will induce a depressive-like phenotype.
3. Studies which use an outcome measure to assess animals who receive a drug intervention. Once an outcome measure has been established (known to be valid and reliable), experimenters know an outcome measure can reliably measure a depressive-like phenotype.

As a starting point, in order to be thorough, we will look at experiments that investigate the differences between a control group and a lesion or model group. This will provide a basis for characterising a depression model. In these experiments, investigators are directly manipulating a variable intended to produce a depressive-like phenotype and measuring the effects of this manipulation on a given outcome measure. These experiments may or may not include the presence of a drug group/arm.

On this basis, we will characterize all the known models/lesions. From this, drug interventions that have been tested on known models can be characterized. Secondly, the known outcome measures will be extracted from the control vs model investigations. Once the most commonly used outcome measures are known, there is further scope for characterizing the studies that investigate drug interventions with known outcome measures. We aim to unpick these different experiment design types and evaluate the evidence from all of these (Table 1).

TABLE 1 Stage of the project at time of protocol submission

Stage of process	Started	Completed
Preliminary searches	Yes	Yes
Piloting study selection	Yes	Yes
Formal screening with final search criteria	Yes	Yes
Data extraction from included papers	No	No
Quality assessment	No	No
Data analysis	No	No
Manuscript writing	No	No

2 | OBJECTIVES OF THIS SYSTEMATIC REVIEW

2.1 | Specify the disease/health problem of interest

We will investigate how depression is modelled *in vivo*. Depression is defined as specified in the DSM-IV-R under the clinical diagnoses of "Depression," "Major Depressive Disorder (MDD) (Single Episode or Recurring)," "Dysthymia (Persistent Depressive Disorder)," "Depressive Disorder Due to General Medical Condition," "Other Specified Depressive Disorder" and/or "Unspecified Depressive Disorder." This includes depression at any life stage in any gender. In preclinical modelling of depression, some methods induce depressive-like behaviour as a single manifestation rather than modelling the full range of features associated with the clinical diagnoses. Therefore, any model that attempts to mimic one or several major symptoms of depression in an animal model will be considered.

Not all aspects of the human condition can be modelled. Some typically modelled phenotypes in non-human animals include anhedonia and disturbances in sleep and/or food consumption. However, we will not exclude the possibility that novel research has attempted to investigate other aspects not previously modelled in non-human animals, and therefore, there are no restrictions on what phenotypes are modelled, only that they are present in the manifestation of the condition in humans.

2.2 | Specify the population/species studied

All preclinical studies on any animal species at any stage of development will be included.

2.3 | Specify the intervention/exposure

This study will investigate any mode of inducing depressive behaviour or a model that seeks to mimic the human condition or symptoms of depression using genetic, surgical, pharmacological, developmental or behavioural interventions or a combination of interventions. We will include models induced acutely, chronically, genetically or through a combination of these methods. We will also consider experiments where the efficacy of a treatment or intervention is tested in such models.

2.4 | Specify the control population

Studies will be included in this review if they include a suitable control, defined as a cohort of animals that have not been exposed to the method of inducing depressive-like behaviour that was used to create the depressive model. The control cohort may have received an appropriate equivalent, for example, sham surgery instead of lesion or placebo without the active ingredient. For studies that investigate treatment efficacy, a suitable control is defined as a cohort of animals that have had the same exposure to model the disorder as those that are given a treatment but has not been exposed to the treatment tested and may instead receive a placebo in an equivalent route of administration. For studies investigating drug intervention on an outcome measure, a suitable control is defined as a cohort of animals that are not exposed to the drug treatment and may instead receive a placebo in an equivalent route of administration.

2.5 | Specify the outcome measures

2.5.1 | Primary outcome measure

The primary outcome measure is behavioural outcome measures of animal studies inducing depressive-like phenotype.

2.5.2 | Secondary outcome measures

Secondary outcomes include anatomical outcomes, electrophysiological outcomes, neurochemical outcomes and prevalence of reporting of measures to reduce risk of bias.

2.5.3 | Tertiary outcome measures

Tertiary outcomes include drug efficacy; inter-rater agreement in the application of the inclusion criteria; and sample size, genomic, proteomic and metabolomic outcomes.

2.6 | Research Questions

1. How are animal models of depression induced?
2. What type of outcome measures are assessed in animal models of depression?
3. How precise and accurate are the outcome measures at assessing induced behaviours?
4. To what extent are the outcomes measured in animal models relevant to the endpoints investigated in human trials?
5. How efficacious are different drug interventions in reducing observed manifestations in *in vivo* animal models?

3 | METHODS

3.1 | Search and study identification

3.1.1 | Identify literature databases to search

Both PubMed and Embase will be searched.

3.1.2 | Define electronic search strategies

See attached for PubMed search terms (Appendix S1, Supporting Information) and Embase search terms (Appendix S2). The animal search filter used for both PubMed and Embase search strings was developed by the Systematic Review Centre for Laboratory Animal Experimentation, Radboud University Medical Centre (SYRCLE). Reference: A step-by-step guide to systematically identify all relevant animal studies. Marlies Leenaars, Carlijn R Hooijmans, Nieky van Veggel, Gerben ter Riet, Mariska Leeflang, Lotty Hooft, Gert Jan van der Wilt, Alice Tillema, Merel Ritskes-Hoitinga. Laboratory Animals. Vol 46, Issue 1, pp. 24 - 31. First published date: January-01-2012. 10.1258/la.2011.011087

3.1.3 | Identify other sources for study identification

Relevant recent reviews will be identified via an additional PubMed search, and the reference list will be searched for any primary research articles that were not identified with the search.

3.2 | Study selection procedure

3.2.1 | Define screening phases (e.g. pre-screening based on title/abstract, full-text screening, both)

PubMed and Embase search results will be downloaded to EndNote or Reference Manager 12; duplicates will be removed and the full text of articles retrieved where available using the inbuilt feature.

Screening Phase 1: Title and abstracts retrieved from PubMed and Embase will be screened.

Screening Phase 2: Full-text papers will be screened concurrently with data extraction.

3.2.2 | Specify number of observers per screening phase: method of screening

3.2.2.1 | Phase 1

A machine learning approach is proposed to assist with the screening phase for inclusion and exclusion criteria. A seed set of papers will be screened by 2 independent human screeners upon which the machine learning algorithm can be trained. Any discrepancies will be resolved by a third human screener. We will pilot the most promising approaches in the context of an ongoing collaboration where we are developing machine learning tools for systematic review. The protocol for this approach is under development and will be uploaded to the CAMARADES website (camarades.info).

Pilot testing of the machine learning algorithm: A random sample of 2000 papers from the overall 70 365 studies identified with the search string was screened for suitability of inclusion by 2 independent reviewers. The decision from the 2000 papers (included or excluded) were used, along with title and abstract, as training sets for several different machine learning algorithms developed by collaborators in the Systematic Living Information Machine (SLIM) consortium. Based on the preliminary sensitivity and specificity analyses and data from a similar project testing the same algorithms using a neuropathic pain dataset,[16,17] an estimated Work Saved over Sampling at 95% recall level (WSS@95% = (TN + FN/N) − 0.5) of above 50% will easily be achieved. Using this approach can reduce the screening workload by at least an estimated 50%, reducing the number of papers needed to be screened by 2 independent human reviewers to less than 35 183 papers.

Quality assessment: A small sub-section of the papers, included and excluded papers, that the machine learning algorithm classifies will be checked by a human screener to ensure the performance of the algorithm.

Validation: We will validate the machine learning techniques for screening by sampling; as opposed to having 2 human screeners manually screen every record. A randomly selected proportion of records that are included and excluded by the algorithm will be double checked by human screeners to ensure the gold standard is maintained. We will continuously monitor the articles screened by the machine learning algorithm by sampling. The machine learning approaches must reach comparable levels to human screening gold standard of at least 95% sensitivity, after which the machine learning algorithm that is maximized for specificity will be chosen.

3.2.2.2 | Phase 2

Two independent screeners are responsible for full-text analysis and data extraction, with the aid of machine learning and text mining tools where appropriate, for example, risk of bias classification. A third independent screener will resolve any discrepancies.

3.3 | Study selection criteria

3.3.1 | Type of study design

3.3.1.1 | Inclusion criteria

Any article providing primary data of an animal model of depression or depressive-like phenotype with an appropriate control group (specified above).

3.3.1.2 | Exclusion criteria

Review article, editorials, case reports, letters or comments, conference or seminar abstracts, studies providing primary data but not appropriate control group.

3.3.2 | Type of animals/population

3.3.2.1 | Inclusion criteria

Animals of all ages, sexes and species, where depression-like phenotype intended to mimic the human condition have been induced. Including animal models where depressive-like phenotypes are induced in the presence of a comorbidity (e.g. obesity or cancer).

3.3.2.2 | Exclusion criteria

Human studies and *ex vivo*, *in vitro* or *in silico* studies. Studies will be excluded if authors state an intention to induce or investigate only anxiety or anxious behaviour. Studies will be excluded if there is no experimental intervention on the animals (e.g. purely observational studies).

3.3.3 | Type of intervention

3.3.3.1 | Inclusion criteria

All studies that claim to model depression or depressive-like phenotypes in animals. Studies that induce depressive behaviour or model depression and that also test a treatment or intervention (prior or subsequent to model induction), with no exclusion criteria based on dosage, timing or frequency.

3.3.3.2 | Exclusion criteria

Studies that investigate treatments or interventions, but no depressive behaviour or model of depression is induced (e.g. toxicity and side-effect studies).

3.3.4 | Outcome measures

3.3.4.1 | Inclusion criteria

Studies measuring behavioural, anatomical and structural, electrophysiological, histological and/or neurochemical outcomes and where genomic, proteomic or metabolomic outcomes are measured in addition to behavioural, anatomical, electrophysiological, histological or neurochemical outcomes.

3.3.4.2 | Exclusion criteria

Where metabolic outcome measures are the primary outcome measure of a study. Where genomic, proteomic, metabolic or metabolomic outcomes are the sole outcome measures in a study, they will be excluded.

3.3.5 | Language restrictions

3.3.5.1 | Inclusion criteria

All languages (using automated translations where required).

3.3.5.2 | Exclusion criteria

None.

3.3.6 | Publication date restrictions

3.3.6.1 | Inclusion criteria

All publication dates.

3.3.6.2 | Exclusion criteria

None.

3.3.7 | Other

3.3.7.1 | Inclusion criteria

Studies must investigate methods or models that induce depressive phenotype/s *in vivo*, or authors must claim that they investigate a model of depression.

3.3.7.2 | Exclusion criteria

Studies claiming to induce only anxiety behaviour or a model of anxiety. In cases where both models of anxiety and depression are investigated, the study will be included, and only the depression-related data will be extracted. In the case of data duplication (2 or more papers reporting the same data), the paper reporting the smallest dataset or fewest outcomes will be excluded. Studies will be excluded if they model aspects of bipolar disorder, manic symptoms, obsessive-compulsive behaviours, panic disorder or psychotic symptoms.

3.3.8 | Order of priority exclusion criteria per screening phase

Selection phase 1: screening based on title and abstract

1. Article must be primary research article (excluding reviews, comments or letters).
2. Exclude studies on humans.
3. Exclude *ex vivo*, *in vitro* or *in silico* investigations.
4. Exclude study if no depressive behaviour or model of depression has been induced.

Selection phase 2: full text screening

1. The above criteria from selection phase 1.
2. Exclude if no appropriate outcome is measured.
3. Exclude if no appropriate control group.
4. Where sufficient data cannot be extracted and authors do not respond to requests for required information.
5. Exclude the study with the least information in the case of multiple publications describing the same work.[18]

4 | STUDY CHARACTERISTICS TO BE EXTRACTED

4.1 | Study meta-data

The first author, corresponding author, year, title, journal name, source of funding and DOI will be extracted.

4.2 | Study design characteristics

The number of animals in the experimental and control groups will be extracted. If the number of animals is given as a range, the most conservative estimate will be extracted. The category of experimental design of the study will be extracted (1. Control vs model induction, 2. Model vs drug, 3. Testing drug on outcome measure).

4.3 | Animal model characteristics

The species, strain, sex (male or female), age and/or weight of animal will be extracted.

4.4 | Method of model induction/intervention characteristics

For studies that induce a model of depression, the method used to induce depressive-like behaviour will be extracted as well as the duration of the model induction. The category of type of model induction will be extracted (i.e., genetic, surgical, pharmacological, developmental or behavioural model induction or a combination of interventions). We will extract information about whether models were induced acutely or chronically or both. If applicable, the following information will be extracted for the method of model induction: the dosage of intervention given, route of delivery, mode of delivery and how long the intervention was given for. The length of time between model induction and outcome measurement will be extracted as well as the length of time between the model induction, any treatment and outcome measurement. For studies with several methods of model induction given, data from all time points will be extracted along with details of each method of model induction.

The information regarding outcome measures described below will be extracted both from experiments describing the induction of a model of depression and from those describing the efficacy of drugs in such models.

4.4.1 | Outcome measures

1. Summary outcome data for each group (mean), including whether variance is reported as SD or SEM and the number of animals per group.
2. Details of the outcome measure (e.g. the sub-type or name of the outcome measure and, e.g. in the case of food restriction, the length of time the animal was restricted for).
3. The number of times the outcome measure was assessed.
4. The number of different outcome measures the animal was tested on.
5. The category of the behaviour or biomarker the outcome measure is measuring (e.g. anhedonia, sleep or weight loss, markers of oxidative stress)
6. Any measures taken before the disease model induction will be extracted. The details of the before-and-after comparison will be extracted.
7. Has an appropriate outcome measure been selected for use? Studies that induce a depression model and investigate the effect of a subsequent drug intervention should select a suitable test to measure an outcome (e.g. an outcome measure that does not rely on the same mechanism/behaviour as behaviour that might be affected by side-effects of a given drug).

4.4.2 | Treatment characteristics

The following information regarding the treatments tested will be extracted: the dosage of treatment given, route of delivery, mode of delivery, how long the treatment was given for. The length of time between the administration of the treatment and outcome measurement will be extracted as well as the length of time between the model induction and any treatment given if applicable. This information will be extracted regardless of whether an experiment simply assesses an outcome measure for a given drug treatment or if an experiment has induced a model of depression and tests a drug treatment.

4.4.3 | Other

The number of excluded animals will be extracted, and the reason for their exclusion, if reported, will be extracted.

5 | RISK OF BIAS & STUDY QUALITY

5.1 | Criteria to assess the internal validity of included studies

An adjusted CAMARADES checklist will be used to assess risk of bias, including the following criteria:

1. Publication in a peer-reviewed journal.
2. Reporting of random allocation.
3. Reporting of blinding of the conduct of the experiment.
4. Reporting of blinded assessment of outcome.
5. Use of comorbid animals (refers to animals where depression is investigated in the presence of another medical condition, e.g. stroke or diabetes).
6. Reporting of a sample size calculation.

7. Reporting of compliance with animal welfare regulations.
8. Reporting of a potential conflict of interest.
9. Reporting of exclusions of animals.
10. Whether a study protocol is available dated before the experiments began.[19]

We will report the median number of study checklist items scored and the interquartile range.

6 | COLLECTION OF OUTCOME DATA

6.1 | Methods for data extraction/retrieval

1. Numerical data will be extracted from the full text of publications (mean, SD or SEM, *P* values (exact *P* -value where possible) and group sample size).
2. In studies where data are presented only graphically, the software Universal Desktop Ruler, or a similar tool, will be used to extract the data into numerical values. For certain PDF presentations, it may be possible to use data mining approaches to extract these data.
3. If any data are missing, the corresponding authors will be contacted.
4. In the absence of a response from authors (we will allow 2 months to reply with a follow-up email sent after the 1st month), data will be excluded from analysis.

If the screeners or extractors consider that 2 sources may describe the same data, we will contact the authors seeking clarification. If we receive no response, we will include only the most recent data source.

7 | DATA ANALYSIS/SYNTHESIS

7.1 | Data gathering and combination

All data will be gathered and entered in the CAMARADES-SyRF database. We will provide a qualitative summary along with several separate meta-analyses, where feasible.

7.2 | How the decision as to whether a meta-analysis is appropriate will be made

Based on pilot analyses of a random sample of 2000 from the overall 70 365 studies identified with the search string, approximately 15% of the total records are expected to be relevant to the research question and included in subsequent meta-analyses. This is similar to previous systematic reviews in models of psychiatric disorders conducted at CAMARADES where about 10% to 15% of the studies were included in the analysis. We expect high heterogeneity between studies due to differences in the study designs; therefore, a meta-analysis is proposed to investigate sources of this heterogeneity.

8 | IF A META-ANALYSIS SEEMS FEASIBLE/SENSIBLE

8.1 | Effect measure to be used

Mean, SD or SEM and group sample size will be extracted for all outcome measures for both experimental and control groups to calculate pooled effect size. Where a single control group serves multiple intervention groups, the size of the control group used in the meta-analysis will be adjusted by dividing it by the number of intervention groups it serves. If the number of animals is presented in a range, the most conservative estimate will be extracted (e.g. if presented as n = 6–12, we will consider n = 6). *P* values, exact *P* value where possible, will be extracted from primary analyses between model and control and intervention and model in order to conduct *P*-curve analysis.

Categorical or qualitative information relating to the outcome measures, such as the behavioural measure or the symptom the model is trying to elucidate, will be extracted into a text/comment field or into a form drop-down menu.

A decision will be made once the data has been extracted as to which effect size is the most appropriate to use. As most outcome measures are continuous variables, and outcome measures are not likely to be measured on the same scale, Normalized Mean Difference (NMD) effect sizes will be calculated where possible. This effect size calculation will be used where an appropriate "sham" or "control" group is present[20] or where it is possible to impute the outcome in a "normal" animal. If the data are unsuitable for calculating NMD, Standardized Mean Difference (SMD) will be used. NMD and SMD will be calculated using the equations outlined in Vesterinen and colleagues.[20]

8.2 | Statistical model of analysis

The data extracted will in all likelihood cover different species, ages and sexes, as well as different study designs and models of induction. Therefore, the true effect size is likely to differ between studies, and a random effects model will be used.[21]

Statistical analyses will be performed using Stata, Statistical Software (College Station, TX: StataCorp LP).

8.3 | Statistical methods to assess heterogeneity

Cochran's Q will be used for assessing heterogeneity; Q is used to calculate the excess variance (Q–k, where k is the degrees of freedom). A *P* value can be calculated for Q, giving an indication of whether all studies share a common effect size ($P < 0.05$) or not ($P > 0.05$). I^2 will be used to report heterogeneity as this describes the proportion of observed variance that reflects true differences in effect size between studies.[18]

8.4 | Specify which study characteristics will be examined as potential sources of heterogeneity (subgroup analysis)

Meta-regression will be used to investigate the impact of different study characteristics on the outcome, where the effect estimate

(NMD or SMD) is the dependent variable. Categorical variables will be transformed into dummy variables. Where there are sufficient data, a multivariate meta-regression will be used for both model induction and drug models. At least 10 independent comparisons per covariate investigated are required.[21] If there are insufficient data for multivariate meta-regression, univariate analysis will be used, requiring a total of at least 25 independent comparisons.

8.5 | Model induction model

Sub-groups analyses:

1. Species of animals (mice vs rats vs etc. vs all)
2. Sex of animals (male vs female vs mixed)
3. Type of animal model
4. Method of model induction (e.g. developmental, genetic, pharmacological, lesion or combination)
5. The outcome measure(s) investigated (behavioural, electrophysiological, neurochemical, anatomical)
6. Number of times the outcome assessment was measured (once vs several)
7. The time from model induction to time of outcome assessment
8. Randomisation (yes/no)
9. Blinding:
   a. Allocation concealment (yes/no)
   b. Assessment of outcome (yes/no)
10. Source of funding (public vs industry)

A separate model will be used to investigate the effect of drug intervention on outcomes.

8.6 | Drug model

Sub-group analyses:

1. Drug Treatment or Intervention.
2. Method of model induction (e.g. developmental, genetic, pharmacological, lesion or combination), if applicable.
3. The outcome measure(s) investigated (behavioural, electrophysiological, neurochemical, anatomical).
4. Treatment or intervention dose.
5. Treatment or intervention route.
6. Number of times the treatment or intervention is administered.
7. Time the treatment is given in relation to model induction (investigated separately per treatment, pre- or post-model induction).
8. Time the treatment is given in relation to time of outcome assessment.
9. Randomisation (yes/no).
10. Blinding:
    a. Allocation concealment (yes/no)
    b. Assessment of outcome (yes/no)
11. Source of funding (public vs industry)

Sensitivity analyses will be performed to assess how missing data from study characteristics and effect size might have affected the results. This will be presented in the form of a summary table.

8.7 | Correction for multiplicity of testing

Where there are more than 2 groups being compared in a univariate model, we will use the Holm-Bonferroni correction for multiplicity of testing.

8.8 | Method for assessment of risk of publication bias

Risk of publication bias analyses will be assessed using funnel plot assessment, *P*-curve analysis and Egger's regression. Trim and fill analysis will be used to identify potentially missing studies. Analyses will be carried out using SigmaPlot and STATA software package (StataCorp LP; SYSTAT Software Inc).

ACKNOWLEDGEMENTS

This research is, in part, supported by a scholarship from the Aarhus-Edinburgh Excellence in European Doctoral Education Project. This research is, in part, funded by a grant from the National Centre for the Replacement, Refinement & Reduction of Animals in Research (NC/L000970/1). We thank members from the SLIM (Systematic Living Information Machine) consortium for their collaboration. We thank Zsanett Bahor, Sarah McCann and Hanna Vesterinen for their comments on versions of the protocol.

Conflict of interest

M.M. is co-editor-in-chief of EBPM.

REFERENCES

1. National Institute for Health and Clinical Excellence (Great Britain). *Depression in Adults: Recognition and Management*. London, UK: National Institute for Health and Clinical Excellence; 2009.
2. Marcus M, Yasamy MT, van Ommeren M, Chisholm D, Saxena S. Depression: a global public health concern. World Health Organisation; 2012. Retrieved November 5, 2015.
3. Gustavsson A, Svensson M, Jacobi F, et al. Cost of disorders of the brain in Europe 2010. *Eur Neuropsychopharmacol*. 2011;21(10):718-779.
4. World Health Organisation (WHO), The World Bank. Out of the shadows: making mental health a global development priority. World Bank Group/IMF Spring Meeting, Washington DC, April 2016.
5. Geddes JR, Carney SM, Davies C, et al. Relapse prevention with antidepressant drug treatment in depressive disorders: a systematic review. *Lancet (Engl)*. 2003;361(9358):653-661.
6. Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am J Psychiatry*. 2006;163(11):1905-1917.
7. Nestler EJ, Gould E, Manji H. Preclinical models: status of basic research in depression. *Biol Psychiatry*. 2002;52(6):503-528.
8. Slattery DA, Cryan JF. The ups and downs of modelling mood disorders in rodents. *ILAR J*. 2014;55(2):297-309.
9. McCann SK, Irvine C, Mead GE, et al. Efficacy of antidepressants in animal models of ischemic stroke: a systematic review and meta-analysis. *Stroke*. 2014;45(10):3055-3063.

10. Mead G, Hackett ML, Lundstrom E, Murray V, Hankey GJ, Dennis M. The FOCUS, AFFINITY and EFFECTS trials studying the effect(s) of fluoxetine in patients with a recent stroke: a study protocol for three multicentre randomised controlled trials. *Trials*. 2015;16:369.
11. Krishnan V, Nestler EJ. The molecular neurobiology of depression. *Nature*. 2008;455(7215):894-902.
12. Kieburtz K, Olanow CW. Translational experimental therapeutics: the translation of laboratory-based discovery into disease-related therapy. *Mt Sinai J Med*. 2007;74(1):7-14.
13. MQ: Transforming Mental Health. Depression: asking the right questions; 2016. http://b.3cdn.net/joinmq/10fbab8ed26626f32e_a3m6bjx2e.pdf. Accessed January 01, 2016.
14. de Vries RB, Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. A search filter for increasing the retrieval of animal studies in Embase. *Lab Anim*. 2011;45(4):268-270.
15. Geyer MA, Markou A. Animal models of psychiatric disorders. In: Bloom FE, Kupfer D, eds. *Psychopharmacology: fourth generation of progress*. New York, NY: Raven; 1995:787-798.
16. Currie GL, Sherratt N, Colvin LA, et al. Search for studies using animal models of neuropathic pain [Data set]. *Zenodo*. 2015. https://doi.org/10.5281/zenodo.35448.
17. Howard BE, Phillips J, Miller K, et al. SWIFT-Review: a text-mining workbench for systematic review. Systematic Reviews. 2016;5:87. doi:10.1186/s13643-016-0263-z.
18. Bahor Z, Macleod M, Thompson L, Sena S, Nunes-Fonseca C.. Improving our understanding of the in vivo modelling of psychotic disorders; 2014. http://www.dcn.ed.ac.uk/camarades/files/Updated%20Protocol.pdf
19. Zwetsloot PP, Jansen LSJ, Végh AMD, et al. Cardiac stem cell treatment in myocardial infarction: protocol for a systematic review and meta-analysis of preclinical studies. *Evid Based Preclin Med*. 2015;2(1):10-15.
20. Vesterinen HMES, Sena KJ, Egan TC, Hirst L, Churolov GL, Currie A, Antonic DW, Howells, and Currie MR Macleod. Meta-analysis of data from animal studies: a practical guide." Journal of neuroscience methods 221 (2014): 92-102.
21. Borenstein M, Hedges LV, Higgins J, Rothstein HR. *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, Ltd; 2009.
22. de Vries RB, Wever KE, Avey MT, Stephens ML, Sena ES, Leenaars M. The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *ILAR J*. 2014;55(3):427-437.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Bannach-Brown A, Liao J, Wegener G, and Macleod M, Understanding *in vivo* modelling of depression in non-human animals: a systematic review protocol. *Evidence-based Preclinical Medicine*. 2017;e00024. https://doi.org/10.1002/ebm2.24

Appendix 1: PubMed Search for "Understanding *In Vivo* Modelling of Depression: A Systematic Review Protocol"

((((((("depressive disorder"[MeSH Terms] OR ("depressive"[All Fields] AND "disorder"[All Fields]) OR "depressive disorder"[All Fields] OR "depression"[All Fields] OR "depression"[MeSH Terms]) OR depressive[All Fields] OR ("depressive behavior"[All Fields] OR "depressive behaviors"[All Fields] OR "depressive behavior"[All Fields] OR "depressive behaviours"[All Fields]) OR ("dysthymic disorder"[MeSH Terms] OR ("dysthymic"[All Fields] AND "disorder"[All Fields]) OR "dysthymic disorder"[All Fields] OR "dysthymia"[All Fields])) NOT (("stress disorders, post-traumatic"[MeSH Terms] OR ("stress"[All Fields] AND "disorders"[All Fields] AND "post-traumatic"[All Fields]) OR "post-traumatic stress disorders"[All Fields] OR "ptsd"[All Fields]) NOT ("depressive disorder"[MeSH Terms] OR "depression"[MeSH Terms]))) NOT (("postpartum period"[MeSH Terms] OR ("postpartum"[All Fields] AND "period"[All Fields]) OR "postpartum period"[All Fields] OR "postpartum"[All Fields]) NOT ("depressive disorder"[MeSH Terms] OR "depression"[MeSH Terms]))) NOT ("schizophrenia"[MeSH Terms] NOT ("depressive disorder"[MeSH Terms] OR "depression"[MeSH Terms])))

AND

(("animal experimentation"[MeSH Terms] OR "models, animal"[MeSH Terms] OR "invertebrates"[MeSH Terms] OR "Animals"[Mesh:noexp] OR "animal population groups"[MeSH Terms] OR "chordata"[MeSH Terms:noexp] OR "chordata, nonvertebrate"[MeSH Terms] OR "vertebrates"[MeSH Terms:noexp] OR "amphibians"[MeSH Terms] OR "birds"[MeSH Terms] OR "fishes"[MeSH Terms] OR "reptiles"[MeSH Terms] OR "mammals"[MeSH Terms:noexp] OR "primates"[MeSH Terms:noexp] OR "artiodactyla"[MeSH Terms] OR "carnivora"[MeSH Terms] OR "cetacea"[MeSH Terms] OR "chiroptera"[MeSH Terms] OR "elephants"[MeSH Terms] OR "hyraxes"[MeSH Terms] OR "insectivora"[MeSH Terms] OR "lagomorpha"[MeSH Terms] OR "marsupialia"[MeSH Terms] OR "monotremata"[MeSH Terms] OR "perissodactyla"[MeSH Terms] OR "rodentia"[MeSH Terms] OR "scandentia"[MeSH Terms] OR "sirenia"[MeSH Terms] OR "xenarthra"[MeSH Terms] OR "haplorhini"[MeSH Terms:noexp] OR "strepsirhini"[MeSH Terms] OR "platyrrhini"[MeSH Terms] OR "tarsii"[MeSH Terms] OR "catarrhini"[MeSH Terms:noexp] OR "cercopithecidae"[MeSH Terms] OR "hylobatidae"[MeSH Terms] OR "hominidae"[MeSH Terms:noexp] OR "gorilla gorilla"[MeSH Terms] OR "pan paniscus"[MeSH Terms] OR "pan troglodytes"[MeSH Terms] OR "pongo pygmaeus"[MeSH Terms]) OR ((animals[tiab] OR animal[tiab] OR mice[Tiab] OR mus[Tiab] OR mouse[Tiab] OR murine[Tiab] OR woodmouse[tiab] OR rats[Tiab] OR rat[Tiab] OR murinae[Tiab] OR muridae[Tiab] OR cottonrat[tiab] OR cottonrats[tiab] OR hamster[tiab] OR hamsters[tiab] OR cricetinae[tiab] OR rodentia[Tiab] OR rodent[Tiab] OR rodents[Tiab] OR pigs[Tiab] OR pig[Tiab] OR swine[tiab] OR swines[tiab] OR piglets[tiab] OR piglet[tiab] OR boar[tiab] OR boars[tiab] OR "sus scrofa"[tiab] OR ferrets[tiab] OR ferret[tiab] OR polecat[tiab] OR polecats[tiab] OR "mustela putorius"[tiab] OR "guinea pigs"[Tiab] OR "guinea pig"[Tiab] OR cavia[Tiab] OR callithrix[Tiab] OR marmoset[Tiab] OR marmosets[Tiab] OR cebuella[Tiab] OR hapale[Tiab] OR octodon[Tiab] OR chinchilla[Tiab] OR chinchillas[Tiab] OR gerbillinae[Tiab] OR gerbil[Tiab] OR gerbils[Tiab] OR jird[Tiab] OR jirds[Tiab] OR merione[Tiab] OR meriones[Tiab] OR rabbits[Tiab] OR rabbit[Tiab] OR hares[Tiab] OR hare[Tiab] OR diptera[Tiab] OR flies[Tiab] OR fly[Tiab] OR dipteral[Tiab] OR drosphila[Tiab] OR drosophilidae[Tiab] OR cats[Tiab] OR cat[Tiab] OR carus[Tiab] OR felis[Tiab] OR nematoda[Tiab] OR nematode[Tiab] OR nematoda[Tiab] OR nematode[Tiab] OR nematodes[Tiab] OR sipunculida[Tiab] OR dogs[Tiab] OR dog[Tiab] OR canine[Tiab] OR canines[Tiab] OR canis[Tiab] OR sheep[Tiab] OR sheeps[Tiab] OR mouflon[Tiab] OR mouflons[Tiab] OR ovis[Tiab] OR goats[Tiab] OR goat[Tiab] OR capra[Tiab] OR capras[Tiab] OR rupicapra[Tiab] OR chamois[Tiab] OR haplorhini[Tiab] OR

monkey[Tiab] OR monkeys[Tiab] OR anthropoidea[Tiab] OR anthropoids[Tiab] OR saguinus[Tiab] OR tamarin[Tiab] OR tamarins[Tiab] OR leontopithecus[Tiab] OR hominidae[Tiab] OR ape[Tiab] OR apes[Tiab] OR pan[Tiab] OR paniscus[Tiab] OR "pan paniscus"[Tiab] OR bonobo[Tiab] OR bonobos[Tiab] OR troglodytes[Tiab] OR "pan troglodytes"[Tiab] OR gibbon[Tiab] OR gibbons[Tiab] OR siamang[Tiab] OR siamangs[Tiab] OR nomascus[Tiab] OR symphalangus[Tiab] OR chimpanzee[Tiab] OR chimpanzees[Tiab] OR prosimians[Tiab] OR "bush baby"[Tiab] OR prosimian[Tiab] OR bush babies[Tiab] OR galagos[Tiab] OR galago[Tiab] OR pongidae[Tiab] OR gorilla[Tiab] OR gorillas[Tiab] OR pongo[Tiab] OR pygmaeus[Tiab] OR "pongo pygmaeus"[Tiab] OR orangutans[Tiab] OR pygmaeus[Tiab] OR lemur[Tiab] OR lemurs[Tiab] OR lemuridae[Tiab] OR horse[Tiab] OR horses[Tiab] OR pongo[Tiab] OR equus[Tiab] OR cow[Tiab] OR calf[Tiab] OR bull[Tiab] OR chicken[Tiab] OR chickens[Tiab] OR gallus[Tiab] OR quail[Tiab] OR bird[Tiab] OR birds[Tiab] OR quails[Tiab] OR poultry[Tiab] OR poultries[Tiab] OR fowl[Tiab] OR fowls[Tiab] OR reptile[Tiab] OR reptilia[Tiab] OR reptiles[Tiab] OR snakes[Tiab] OR snake[Tiab] OR lizard[Tiab] OR lizards[Tiab] OR alligator[Tiab] OR alligators[Tiab] OR crocodile[Tiab] OR crocodiles[Tiab] OR turtle[Tiab] OR turtles[Tiab] OR amphibian[Tiab] OR amphibians[Tiab] OR amphibia[Tiab] OR frog[Tiab] OR frogs[Tiab] OR bombina[Tiab] OR salientia[Tiab] OR toad[Tiab] OR toads[Tiab] OR "epidalea calamita"[Tiab] OR salamander[Tiab] OR salamanders[Tiab] OR eel[Tiab] OR eels[Tiab] OR fish[Tiab] OR fishes[Tiab] OR pisces[Tiab] OR catfish[Tiab] OR catfishes[Tiab] OR siluriformes[Tiab] OR arius[Tiab] OR heteropneustes[Tiab] OR sheatfish[Tiab] OR perch[Tiab] OR perches[Tiab] OR percidae[Tiab] OR perca[Tiab] OR trout[Tiab] OR trouts[Tiab] OR char[Tiab] OR chars[Tiab] OR salvelinus[Tiab] OR "fathead minnow"[Tiab] OR minnow[Tiab] OR cyprinidae[Tiab] OR carps[Tiab] OR carp[Tiab] OR zebrafish[Tiab] OR zebrafishes[Tiab] OR goldfish[Tiab] OR goldfishes[Tiab] OR guppy[Tiab] OR guppies[Tiab] OR chub[Tiab] OR chubs[Tiab] OR tinca[Tiab] OR barbels[Tiab] OR barbus[Tiab] OR pimephales[Tiab] OR promelas[Tiab] OR "poecilia reticulata"[Tiab] OR mullet[Tiab] OR mullets[Tiab] OR seahorse[Tiab] OR seahorses[Tiab] OR mugil curema[Tiab] OR atlantic cod[Tiab] OR shark[Tiab] OR sharks[Tiab] OR catshark[Tiab] OR anguilla[Tiab] OR salmonid[Tiab] OR salmonids[Tiab] OR whitefish[Tiab] OR whitefishes[Tiab] OR salmon[Tiab] OR salmons[Tiab] OR sole[Tiab] OR solea[Tiab] OR "sea lamprey"[Tiab] OR lamprey[Tiab] OR lampreys[Tiab] OR pumpkinseed[Tiab] OR sunfish[Tiab] OR sunfishes[Tiab] OR tilapia[Tiab] OR tilapias[Tiab] OR turbot[Tiab] OR turbots[Tiab] OR flatfish[Tiab] OR flatfishes[Tiab] OR sciuridae[Tiab] OR squirrel[Tiab] OR squirrels[Tiab] OR chipmunk[Tiab] OR chipmunks[Tiab] OR suslik[Tiab] OR susliks[Tiab] OR vole[Tiab] OR voles[Tiab] OR lemming[Tiab] OR lemmings[Tiab] OR muskrat[Tiab] OR muskrats[Tiab] OR lemmus[Tiab] OR otter[Tiab] OR otters[Tiab] OR marten[Tiab] OR martens[Tiab] OR martes[Tiab] OR weasel[Tiab] OR badger[Tiab] OR badgers[Tiab] OR ermine[Tiab] OR mink[Tiab] OR minks[Tiab] OR sable[Tiab] OR sables[Tiab] OR gulo[Tiab] OR gulos[Tiab] OR wolverine[Tiab] OR wolverines[Tiab] OR minks[Tiab] OR mustela[Tiab] OR llama[Tiab] OR llamas[Tiab] OR alpaca[Tiab] OR alpacas[Tiab] OR camelid[Tiab] OR camelids[Tiab] OR guanaco[Tiab] OR guanacos[Tiab] OR chiroptera[Tiab] OR chiropteras[Tiab] OR bat[Tiab] OR bats[Tiab] OR fox[Tiab] OR foxes[Tiab] OR iguana[Tiab] OR iguanas[Tiab] OR xenopus laevis[Tiab] OR parakeet[Tiab] OR parakeets[Tiab] OR parrot[Tiab] OR parrots[Tiab] OR donkey[Tiab] OR donkeys[Tiab] OR mule[Tiab] OR mules[Tiab] OR zebra[Tiab] OR zebras[Tiab] OR shrew[Tiab] OR shrews[Tiab] OR bison[Tiab] OR bisons[Tiab] OR buffalo[Tiab] OR buffaloes[Tiab] OR deer[Tiab] OR deers[Tiab] OR bear[Tiab] OR bears[Tiab] OR panda[Tiab] OR pandas[Tiab] OR "wild hog"[Tiab] OR "wild boar"[Tiab] OR fitchew[Tiab] OR fitch[Tiab] OR beaver[Tiab] OR beavers[Tiab] OR jerboa[Tiab] OR jerboas[Tiab] OR capybara[Tiab] OR capybaras[Tiab]) NOT medline[subset])))

NOT

(("review"[Publication Type] OR "review literature as topic"[MeSH Terms] OR "review"[All Fields]) OR ("letter"[Publication Type] OR "correspondence as topic"[MeSH Terms] OR "letter"[All Fields]) OR ("comment"[Publication Type] OR "comment"[All Fields]))

Appendix 2: Embase Search Terms for "Understanding *In Vivo* Modelling of Depression: A Systematic Review Protocol"

1. (PTSD.mp. or posttraumatic stress disorder/) not (depression/ or atypical depression/ or dysthymia/ or major depression/ or treatment resistant depression/ or depressive behavio*.mp.)
2. (bipolar disorder/ or bipolar depression/ or bipolar II disorder/ or rapid cycling bipolar disorder/ or bipolar.mp. or bipolar I disorder/ or bipolar mania/) not (depression/ or atypical depression/ or dysthymia/ or major depression/ or treatment resistant depression/ or depressive behavio*.mp.)
3. (schizophrenia assessment/ or residual schizophrenia/ or simple schizophrenia/ or schizophrenia.mp. or paranoid schizophrenia/ or latent schizophrenia/ or disrupted in schizophrenia 1 protein/ or "Schedule for Affective Disorders and Schizophrenia"/ or catatonic schizophrenia/ or schizophrenia/) not (depression/ or atypical depression/ or dysthymia/ or major depression/ or treatment resistant depression/ or depressive behavio*.mp.)
4. exp animal experiment/ or exp animal model/ or exp experimental animal/ or exp transgenic animal/ or exp male animal/ or exp female animal/ or exp juvenile animal/ or animal/ or chordata/ or vertebrate/ or tetrapod/ or exp fish/ or amniote/ or exp amphibia/ or mammal/ or exp reptile/ or exp sauropsid/ or therian/ or exp monotremate/ or placental mammals/ or exp marsupial/ or Euarchontoglires/ or exp Afrotheria/ or exp Boreoeutheria/ or exp Laurasiatheria/ or exp Xenarthra/ or primate/ or exp Dermoptera/ or exp Glires/ or exp Scandentia/ or Haplorhini/ or exp prosimian/ or simian/ or exp tarsiiform/ or Catarrhini/ or exp Platyrrhini/ or ape/ or exp Cercopithecidae/ or hominid/ or exp hylobatidae/ or exp chimpanzee/ or exp gorilla/ or exp orang utan/ or (animal or animals or pisces or fish or fishes or catfish or catfishes or sheatfish or silurus or arius or heteropneustes or clarias or gariepinus or fathead minnow or fathead minnows or pimephales or promelas or cichlidae or trout or trouts or char or chars or salvelinus or salmo or oncorhynchus or guppy or guppies or millionfish or poecilia or goldfish or goldfishes or carassius or auratus or mullet or mullets or mugil or curema or shark or sharks or cod or cods or gadus or morhua or carp or carps or cyprinus or carpio or killifish or eel or eels or anguilla or zander or sander or lucioperca or stizostedion or turbot or turbots or psetta or flatfish or flatfishes or plaice or pleuronectes or platessa or tilapia or tilapias or oreochromis or sarotherodon or common sole or dover sole or solea or zebrafish or zebrafishes or danio or rerio or seabass or dicentrarchus or labrax or morone or lamprey or lampreys or petromyzon or pumpkinseed or pumpkinseeds or lepomis or gibbosus or herring or clupea or harengus or amphibia or amphibian or amphibians or anura or salientia or frog or frogs or rana or toad or toads or bufo or xenopus or laevis or bombina or epidalea or calamita or salamander or salamanders or newt or newts or triturus or reptilia or reptile or reptiles or bearded dragon or pogona or vitticeps or iguana or iguanas or lizard or lizards or anguis fragilis or turtle or turtles or snakes or snake or aves or bird or birds or quail or quails or coturnix or bobwhite or colinus or virginianus or poultry or poultries or fowl or fowls or chicken or chickens or gallus or zebra finch or taeniopygia or guttata or canary or canaries or serinus or canaria or parakeet or parakeets or grasskeet or parrot or parrots or psittacine or psittacines or shelduck or tadorna or goose or geese or branta or leucopsis or woodlark or lullula or flycatcher or ficedula or hypoleuca or dove or doves or geopelia or cuneata or duck or ducks or greylag or graylag or anser or harrier or circus pygargus or red knot or great knot or calidris or canutus or godwit or limosa or lapponica or meleagris or gallopavo or jackdaw or corvus or monedula or ruff or philomachus or pugnax or lapwing or peewit or plover or vanellus or swan or cygnus or columbianus or bewickii or gull or chroicocephalus or ridibundus or albifrons or great tit or parus or aythya or fuligula or streptopelia or risoria or spoonbill or platalea or leucorodia or blackbird or turdus or merula or blue tit or cyanistes or pigeon or pigeons or columba or pintail or anas or starling or sturnus or owl or athene noctua or pochard or ferina or cockatiel or nymphicus or hollandicus or skylark or alauda or tern or sterna or teal or crecca or oystercatcher or haematopus or ostralegus or shrew or shrews or sorex or araneus or crocidura or russula or european mole or talpa or chiroptera or bat or bats or eptesicus or serotinus or myotis or dasycneme or daubentonii or pipistrelle or pipistrellus or cat or cats or felis or catus or feline or dog or dogs or canis or canine or

canines or otter or otters or lutra or badger or badgers or meles or fitchew or fitch or foumart or foulmart or ferrets or ferret or polecat or polecats or mustela or putorius or weasel or weasels or fox or foxes or vulpes or common seal or phoca or vitulina or grey seal or halichoerus or horse or horses or equus or equine or equidae or donkey or donkeys or mule or mules or pig or pigs or swine or swines or hog or hogs or boar or boars or porcine or piglet or piglets or sus or scrofa or llama or llamas or lama or glama or deer or deers or cervus or elaphus or cow or cows or bos taurus or bos indicus or bovine or bull or bulls or cattle or bison or bisons or sheep or sheeps or ovis aries or ovine or lamb or lambs or mouflon or mouflons or goat or goats or capra or caprine or chamois or rupicapra or leporidae or lagomorpha or lagomorph or rabbit or rabbits or oryctolagus or cuniculus or laprine or hares or lepus or rodentia or rodent or rodents or murinae or mouse or mice or mus or musculus or murine or woodmouse or apodemus or rat or rats or rattus or norvegicus or guinea pig or guinea pigs or cavia or porcellus or hamster or hamsters or mesocricetus or cricetulus or cricetus or gerbil or gerbils or jird or jirds or meriones or unguiculatus or jerboa or jerboas or jaculus or chinchilla or chinchillas or beaver or beavers or castor fiber or castor canadensis or sciuridae or squirrel or squirrels or sciurus or chipmunk or chipmunks or marmot or marmots or marmota or suslik or susliks or spermophilus or cynomys or cottonrat or cottonrats or sigmodon or vole or voles or microtus or myodes or glareolus or primate or primates or prosimian or prosimians or lemur or lemurs or lemuridae or loris or bush baby or bush babies or bushbaby or bushbabies or galago or galagos or anthropoidea or anthropoids or simian or simians or monkey or monkeys or marmoset or marmosets or callithrix or cebuella or tamarin or tamarins or saguinus or leontopithecus or squirrel monkey or squirrel monkeys or saimiri or night monkey or night monkeys or owl monkey or owl monkeys or douroucoulis or aotus or spider monkey or spider monkeys or ateles or baboon or baboons or papio or rhesus monkey or macaque or macaca or mulatta or cynomolgus or fascicularis or green monkey or green monkeys or chlorocebus or vervet or vervets or pygerythrus or hominoidea or ape or apes or hylobatidae or gibbon or gibbons or siamang or siamangs or nomascus or symphalangus or hominidae or orangutan or orangutans or pongo or chimpanzee or chimpanzees or pan troglodytes or bonobo or bonobos or pan paniscus or gorilla or gorillas or troglodytes).ti,ab.
5. depression/ or atypical depression/ or dysthymia/ or major depression/ or treatment resistant depression/ or depressive behavio*.mp.
6. (5 not 1 not 2 not 3) and 4
7. comment.pt. or letter.pt. or review.pt
8. ((5 not 1 not 2 not 3) and 4) not 7

2.4.1 Corrigendum to the Protocol

In section 8.3 “Statistical methods to assess heterogeneity” of the published protocol there is an error in the section sentence on the p-value of Q. It should read, “A P-value can be calculated for Q, giving an indication of whether all studies share a common effect size ($P > 0.05$) or not ($P < 0.05$)”.

2.4.2 Implementation of the protocol

The protocol for this study aimed to explicitly state the research questions and aims of this SR and MA, the search strategy used to identify records, the inclusion and exclusion criteria, the key pieces of information to be extracted from the primary articles, and the proposed methods for pooling the information together. A systematic search of PubMed and EMBASE was conducted in May 2016. The search retrieved 70,365 unique research articles that were potentially relevant to the research topic, animal models of depression. With this amount of studies, it was not feasible to screen using two independent reviewers in a reasonable time-frame (estimated 64 person months). Therefore, machine learning algorithms were employed to learn the classification of papers into “Relevant” and “Not Relevant”.

A sub-set of the studies was screened using an online systematic review facility to assist the conduct of systematic reviews, SyRF.org.uk (SyRF). 7000 studies were screened on SyRF by two independent reviewers against the inclusion criteria (see above pre-specified protocol), with a third independent reviewer reconciling any differences. These 7,000 studies were used to train machine learning algorithms to learn the decision-making capacity. The performance was assessed on sensitivity and specificity, and the aim was to achieve performance to pre-specified criteria. The algorithm achieved the desired level of performance and was applied to the remaining unseen documents (63,365). See Chapter 3: Methods Development for the methods and implementation of the machine learning techniques.

The algorithm identified 18,409 documents to be included in the systematic review. The use of text-mining techniques to assist in annotating the included documents reduced the human resources required to categorise documents by topic of interest.

This enabled the next step of systematic review, extracting and analysing data from primary articles. The application of text-mining tools and custom dictionaries is outlined in Chapter 4. The methodology for subsequent systematic reviews and meta-analyses of sub-topics within this broad review has been outlined with a separate protocol for each review.

The following two chapters outline and detail the automation tools developed, tested, and implemented in this systematic review.

3 Methods Development – Machine Learning for Citation Screening

The work in this chapter has been completed with help from internal and external collaborators. James Thomas from EPPI-Centre, University College London and Piotr Przybyła from National Centre for Text-Mining, University of Manchester contributed significantly with testing their machine learning approaches (Methods > Feature Generation & Classifiers). Kaitlyn Hair and Paula Grill from the CAMARADES team assisted with second screening the training set (Methods > Training Set). Special thanks to Jing Liao from CAMARADES for assistance throughout the project. This chapter has been submitted for publication and an earlier version of the manuscript is available as a preprint on BioRxiv (Bannach-Brown et al., 2018).

3.1 Background

The rate of publication of primary research is increasing exponentially within biomedicine (Bornmann & Mutz, 2015). Researchers find it increasingly difficult to keep up with new findings and discoveries even within a single biomedical domain, an issue that has been emerging for a number of years (Cohen et al., 2010). Synthesising research – either informally or through systematic reviews - becomes increasingly resource intensive as searches retrieve larger and larger corpuses of potentially relevant papers for reviewers to screen for relevance to the research question at hand.

This increase in rate of publication is seen in the animal literature. In an update to a systematic review of animal models of neuropathic pain, 11,880 further unique records were retrieved in 2015, to add to 33,184 unique records identified in a search conducted in 2012. In the field of animal models of depression, the number of unique records retrieved from a systematic search increased from 70,365 in May 2016 to 76,679 in August 2017.

The use of text-mining tools and machine learning (ML) algorithms to aid systematic review is becoming an increasingly popular approach to reduce human burden and

monetary resources required and to reduce the time taken to complete such reviews (Howard et al., 2015; Tsafnat et al., 2014; O'Mara-Eves et al., 2015). ML algorithms are primarily employed at the screening stage in the systematic review process. This screening stage involves categorising records identified from the search into 'Relevant' or 'Not-Relevant' to the research question, typically performed by two independent human reviewers with discrepancies reconciled by a third. This decision is typically made on the basis of the title and abstract of an article in the first instance. In previous experience at CAMARADES (Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies), screening a preclinical systematic review with 33,184 unique search results took 9 months, representing (because of dual screening) around 18 person months in total. Based partly on this, we estimate that a systematic review with roughly 10,000 publications retrieved takes a minimum of 40 weeks. In clinical systematic reviews, Borah and colleagues (2017) showed the average clinical systematic review registered on PROSPERO (International Prospective Register of Systematic Reviews) takes an average 67.3 weeks to complete. ML algorithms can be employed to learn this categorisation ability, based on training instances that have been screened by human reviewers (Cohen et al., 2006).

Several applications of ML are possible. The least burdensome is when a review is being updated, where categorisations from the original review are used to train a classifier, which is then applied to new documents identified in the updated search (Cohen et al., 2006; Cohen et al., 2012; Wallace et al., 2012a). When a screening is performed de novo, without such previous collection, humans first categorise an initial set of search returns, which are used to train an ML model. The performance of the model is then tested (either in a validation set or with k fold cross validation); if performance does not meet a required threshold then more records are screened, chosen either through random sampling or, using active learning (Lewis & Gale, 1994). Active learning involves interaction from the human user during the classification process to optimize the machine learning performance, for example, users may classify either of those with highest uncertainty of predictions (Wallace et al., 2010a; Bornmann & Mutz, 2015; Lui et al., 2016) or those most likely to be included (Miwa et al., 2014; Wallace et al., 2012b; Kontonatsios et al., 2017]. Here we use a de novo

search with subsequent training sets identified by random sampling, and we introduce a novel use of machine prediction, in identifying human error in screening decisions.

Machine learning approaches have been evaluated in context of systematic reviews of several medical problems including drug class efficacy assessment (Cohen et al., 2006; Cohen et al., 2012; Lui et al., 2012), genetic associations (Wallace et al., 2012a), public health (Shemilt et al., 2014; Miwa et al., 2014), cost-effectiveness analyses (Wallace et al., 2012a), toxicology (Howard et al., 2015), treatment effectiveness (Wallace et al., 2010b; Rathbone et al., 2015) and nutrition (Wallace et al., 2010b). To the best of our knowledge there have been only two attempts to apply such techniques to reviews of preclinical animal studies (Howard et al., 2015; Liao et al., 2018). These can be broad and shallow reviews or focussed and detailed reviews, and can have varying prevalence of inclusion.

Here we outline the ML approach taken to assist in screening a corpus for a broad and shallow systematic review seeking to summarise studies using non-human animal models of depression, based on a corpus of 70,365 records retrieved from two online biomedical databases. *Here the aim was to identify the amount of training data required for an algorithm to achieve the level of performance of two independent human screeners, so that we might reduce the human resource required.*

Sena and colleagues developed guidelines for the appraisal of systematic reviews of animal studies (2014). These guidelines consider dual extraction by two independent human reviewers as a feature of a high quality review. From a large corpus of reviews conducted by CAMARADES we estimate the inter-screener agreement to be between 95% and 99%. Errors may occur at random (due to fatigue or distraction) or, more consequentially, systematic error, which, if included in a training set, might be propagated into a ML algorithm. Sources of systematic errors with certain types of records are at greater risk of misclassification. To our knowledge the nature of this 5% residual human error in systematic review methodology has not been formally investigated. The training data used for ML categorisation is based on training instances that has been screened by two independent human screeners.

The aim was to explore the use of established ML algorithms as part of a preclinical systematic review framework at the classification stage, to investigate if the ML algorithms could be used to improve the human gold standard by identifying human screening errors and thus improve the overall performance of ML.

3.2 Methods

We applied two independent machine learning approaches to the screening of a large (70,365 records) systematic review. Because we did could not predict how many training instances would be required, first I selected 2000 records at random to provide the first training set. Of these, only 1993 were suitable due to data deposition errors. These were then screened by 2 human reviewers with previous experience with reviews of animal studies, with a third expert reviewer reconciling any differences. The resulting ML algorithms gave a score between 0 and 1. To ensure that the true sensitivity was likely to be 95% or higher we chose as our cut-point the value for which the lower bound of the 95% confidence interval of the observed sensitivity exceeded 95% when applied to the unseen validation dataset. This level of performance was chosen to match the comparable level of gold-standard human reviewing. I then repeated this process adding a further 1000 randomly selected (996 useable) citations to the training set; and then again adding a further 3000 randomly selected (2760 useable) citations to the training set. At each stage, performance of the approaches was assessed on a validation set of unseen documents, using a number of different metrics. Next, the best performing algorithm was used to identify human errors in the training and validation sets by selecting those with the largest discrepancy between the human decision (characterised as 0 for exclude or 1 for include) and the machine prediction (a continuous variable between 0 and 1). Performance of the approaches trained on the full 5749 records is reported here, and of each of the iterations is available in Supplementary Materials 1. The error analysis was assessed on the net reclassification index, and the performance of the ML approach is compared before and after correcting the errors in human screening using AUC.

3.2.1 Step 1: Application of ML tools to screening of a large preclinical systematic review

3.2.1.1 Training Sets

70,365 potentially relevant records were identified from Pubmed and EMBASE. The search strings were composed of the animal filters devised by the Systematic Review Center for Laboratory animal Experimentation (SYRCLE) (de Vries et al., 2014b; Hooijmans et al., 2010), NOT reviews, comments, or letters AND a depression disorder string (for full search strings see Bannach-Brown et al., 2016a and protocol in Chapter 3). The training set and the validation set were chosen at random from the 70,365 by assigning each record a random number using the RAND function in excel and ranking them from smallest to largest. The training set consisted of 5749 records. The validation set consisted of 1251 records. The training set and validation set were screened by two independent human screeners with any discrepancies reconciled by a third independent human screener. The human screening process involved an online tool (app.syrf.org), which randomly presents a reviewer with a record, with the title and abstract displayed. The reviewer makes a decision about the record, included (1) or excluded (0). A second reviewer is also randomly presented with records. If a record receives two 'included' decisions, the screening for this record is considered complete. If reviewer 1 and reviewer 2 disagree, the record gets presented to a third reviewer who makes a decision. The record then has an average inclusion score of 0.666 or 0.333. Any record that has an inclusion score above 0.6 is included, those scoring less than 0.6 are excluded, and screening is considered complete. Datasets are available on Zenodo, as described in "Availability of Data & Materials" below, Performance was assessed at each level on a validation set of unseen records. The training and validation set were selected consecutively from the initial random ordering. For the training set of 5749 records, the validation set was the subsequent

1251 records. This validation set had more than 150 “included” records, which can give reasonably precise 95% confidence intervals for sensitivity and specificity.

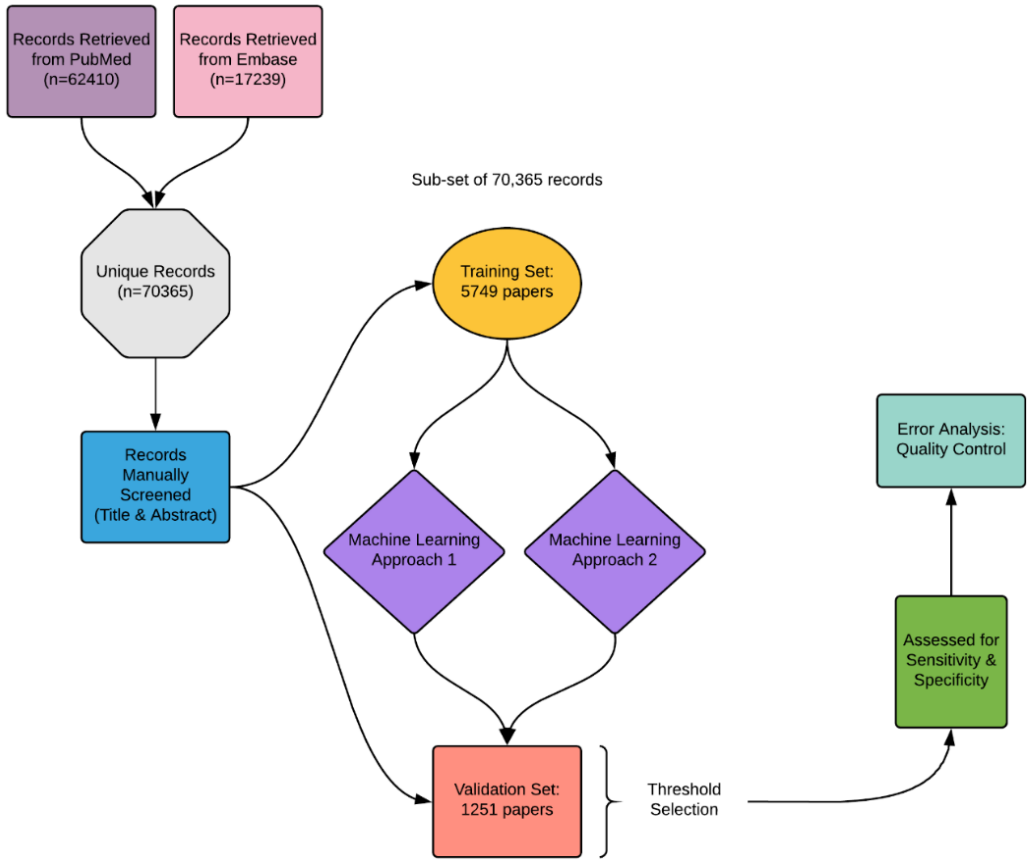


Figure 3.1. Diagram of the Layout of the Study.

3.2.1.2 Feature Generation

First, documents in the training set were transformed into a representation appropriate for the machine learning algorithms. Documents were created by concatenating the title and the abstract. Every case (document) is represented by a fixed number of features, numerical quantities describing certain properties that might be used by the classifier to extract rules and make predictions about inclusion. The classifiers described below used generally similar approaches

We used “bag-of-words” (BoW), a model for representing words in documents to, to characterise document titles and abstracts in both classifiers. To account for the

relative importance of words within a given document, and difference in words used between documents we used 'Term Frequency – Inverse Document Frequency' (TD-IDF). This is defined as:

$$tfidf(w_i, d_j) = tf(w_i, d_j) * \frac{|D|}{|\{d: w_i \in d\}|}$$

The score for the i-th word in context of the j-th document takes into account not only how many times the word occurred there (tf), but also how many other documents (d) from the whole corpus (D) contain it as well. This helps to reduce the score for words that are common for all documents and therefore have little predictive power. This helps the classifier to focus on terms which help to distinguish between documents, rather than on terms which occur frequently (Manning et al., 2008). We allowed n-grams, defined as a string of *n* words derived a given document or article; we did not use stemming; and used the MySQL text indexing functionality "stopword" list to remove frequently occurring words which provide little relevant information for classification purposes (Oracle, 2018).

Because bag-of-words representation generates as many unigram features as there are words in the collection (typically at least several thousand); and many more when using higher-order n-grams, we used additional approaches. Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA) represent textual data in a more efficient way. In LSI (Deerwester et al., 1990), the training set is represented as a matrix, where rows correspond to documents, columns to terms (words or n-grams), while cells contain frequency or TF/IDF score of a given term in a given document. The matrix is then decomposed using a general matrix factorisation technique known as Singular Value Decomposition (SVD) and truncated to the first n dimensions. Because of the properties of SVD the new features will be such linear combinations of features of the old space that minimise the differences between the original and the transformed space. In case of textual data, it means that those words that frequently occur in the same documents (probably because of the similar meaning) will be treated in the same way. The n is set a-priori to a reasonably low value – usually a few hundred. LDA exploits distributional similarities between words but based on explaining document contents using a Bayesian network (Blei et al., 2003). This

method is based on the premise that every document is a mixture of topics, which in turn consist of related words. The correspondence between documents and topics and between topics and words can be inferred via Gibbs sampling process. As a result, similarly to LSI, every document is represented by a sequence of n numbers, indicating how related it is to every topic (Kontonatsios et al., 2017). Unlike in SVD, the model fitness to the data cannot be expressed through the amount of variance of the original matrix it explains and the optimal number of topics may be different for every collection and classification task. Following previous work in the domain (Miwa et al., 2014) and the user guide for MALLET (the tool we use for LDA, which recommends values between 200 and 400) we elected to generate 300 topics Here we use three feature sets, BoW, LDA and SVD (LSI) individually, in pairs and finally all together; preliminary evaluation through the cross-validation on the training set suggests that LDA+SVD and bag-of-words with a simple linear classifier deliver the most robust performance.

3.2.1.4 Classifiers

Following the transformations made in feature selection, the documents are then used to train the machine learning classifier. The classifier most commonly used for document classification in context of systematic reviews (Wallace et al., 2010a; Miwa et al., 2014; Cohen et al., 2012; Wallace et al., 2012a; Lui et al., 2016; Wallace et al., 2012b; Kontonatsios et al., 2017; Wallace et al., 2010b) is the Support Vector Machine (SVM) as it has frequently been used for tasks involving text. SVM is a supervised learning algorithm, learning to classify new documents based on a training set of labelled documents (Mertsalov & McCreary, 2009). This algorithm represents training documents as points in a multi-dimensional space defined by all available features. To be able to classify cases into positive and negative category, it seeks a hyperplane dividing the space into one side corresponding to included documents and the other to excluded ones. Based on the training data, the optimal hyperplane is constructed so that it maximises both the number of training cases located on the “correct” side of decision boundary and their distance from the plane (margin). The new, unseen, documents are then ranked according to their location with respect to the boundary. Those far from it are confidently predicted as included or excluded, according to which side of the plane they lie. The cases which the model has less confidence about will be located close to the hyperplane. Logistic regression is a

similar linear classifier, which instead of hyperplane, seeks such coefficients of a linear combination of feature values that will give high values for positive cases (included documents) and low for negative (excluded documents). Both of these approaches could be enriched with feature selection elements to mitigate the problems with multitude of features.

Three feature sets (BoW, LDA and SVD (LSI)) were tested on SVMs, logistic regression and random forests (Breiman, 2001). The two algorithms described below performed best for this dataset of 70,365 records, on the broad topic of preclinical animal models of depression.

3.2.1.5 Approaches

Here, two approaches were developed independently, using different classification models and feature representations, but sharing the linear classification principles.

3.2.1.5.1 Approach 1

Approach one used a tri-gram 'bag-of-words' model for feature selection and implemented a linear support vector machine with Stochastic Gradient Descent (SGD) as supported by the SciKit-Learn python library (Pedregosa et al., 2011). This classifier was chosen as it is efficient, scales well to large numbers of records, and provides an easily interpretable list of probability estimates when predicting class membership (i.e. scores for each document lying between 0 and 1). Efficiency and interpretability are important, as this classifier is already deployed in a large systematic review platform (Thomas et al. 2010), and any deployed algorithm therefore needs not to be too computationally demanding, and its results understood by users who are not machine learning specialists. The tri-gram feature selection approach without any additional feature engineering also reflects the generalist need of deployment on a platform used in a wide range of reviews: the algorithm needs to be generalisable across disciplines and literatures, and not 'over-fitted' to a specific area. For example, the tri-gram "randomised controlled trial" has quite different implications for classification compared with "randomised controlled trials" (i.e. 'trials' in plural). The former might be a report of a randomised controlled trial; while the latter

is often found in reports of systematic reviews of randomised trials. Stemming would remove the ‘s’ on trials and thus lose this important information. Here, the algorithm needs to be generalisable across disciplines and literatures, and not be ‘over-fitted’ to a specific area. This approach aims to give the best compromise between reliable performance across a wide range of domains and that achievable from a workflow that has been highly tuned to a specific context.

3.2.1.5.2 Approach 2

Approach 2 used a regularised logistic regression model built on LDA and SVD features. Regularisation refers to the mathematical technique of adding a co-efficient to a model to reduce the number of features and reduce over-fitting of the algorithm. Namely, the document text (consisting of title and abstract) was first lemmatised with the tool GENIA tagger (Tsuruoka et al., 205) and then converted into bag of words representation of unigrams, which was then used to create two types of features. First, the word frequencies were converted into a matrix TF/IDF scores, which was then decomposed via SVD implemented in scikit-learn library and truncated to the first 300 dimensions. Second, an LDA model was built using MALLET library (McCallum & Kachites, 2002), setting 300 as a number of topics. As a result each document was represented by 600 features, and an L1-regularised logistic regression model was built using glmnet package (Friedman et al., 2010) in R statistical framework.

In this procedure every document is represented with a constant, manageable number of features, irrespective of corpus or vocabulary size. As a result, we can use a relatively simple classification algorithm and expect good performance with short processing time even for very large collections. This feature is particularly useful when running the procedure numerous times in cross-validation mode for error analysis (see below).

For a given unseen test instance, the logistic regression returns a score corresponding to the probability of it being relevant according to the current model. An optimal cut-off score that gives the best performance is calculated as described above.

3.2.1.6 Assessing Machine Learning Performance

The facets of a machine learning algorithm performance that would be most beneficial to this field of research are high sensitivity (see table 1), at a level comparable to the 95% we estimate is achieved by two independent human screeners. We therefore need to be confident that the sensitivity is 95% or higher, which we do by setting our cut point such that the lower bound of the 95% confidence interval of the observed sensitivity is 95% or higher. Once the level of sensitivity has been reached, the aim is to maximise specificity, to reduce the number of irrelevant records included by an algorithm. Although specificity at 95% sensitivity is the goal, I provide values of other measures for better illustration of the performance.

3.2.1.6.1 Performance metrics

Performance was assessed using sensitivity (or recall), specificity, precision, accuracy, and Work Saved over Sampling (WSS) (see table 1), carried out in R (R version 3.4.2) using the 'caret' package (Kuhn, 2017). 95% Confidence Intervals were calculated using the efficient-score method (Newcombe, 1998). Cut-offs for were determined manually for each approach by taking the score that achieved 95% sensitivity (including the lower 95% confidence level), and the specificity at this score was calculated.

Table 3.1 Equations used to assess performance of machine learning algorithms

Sensitivity or Recall	TP / (TP+FN)
Specificity	TN / (TN+FP)
Precision	TP / (TP+FP)
Accuracy	(TP+TN) / (TP+FP+FN+TN)
WSS@95%	((TN+FN) / N) – (1.0 – 0.95)

All equations from [5].

3.2.2 Step 2: Application of ML tools to training datasets to identify human error

3.2.2.1 Error Analysis Methods

The methodology for the error analysis was outlined in an *a priori* protocol, published on the CAMARADES website on 18th December 2016 (Bannach-Brown et al., 2016b). To generate the machine learning scores for the set of records that were originally used to train the machine (5749 records), the non-exhaustive cross-validation method, 5-fold validation, was used. This method involved randomly partitioning the set of records into 5 equal sized subsamples. One subsample was set aside, and the remaining 4 subsamples were used to train the algorithm (Rodriguez et al., 2010). Thanks to this process, every record has a score computed by a machine learning model built without including it in the training portion. These scores were used to highlight discrepancies or disagreements between machine decision and human decision. The documents were ordered by the machine assigned labels in order of predictive probability, from most likely to be relevant to least likely to be relevant. The original human assigned scores were placed next to the machine-assigned scores, to highlight potential errors in the human decision. A single human reviewer (experienced in animal systematic reviews) manually reassessed the records where discrepancies were highlighted starting with the most discrepant. To avoid reassessing the full 5749 record dataset, a stopping rule was established such that if the initial human decision was correct for five consecutive records, further records were not reassessed.

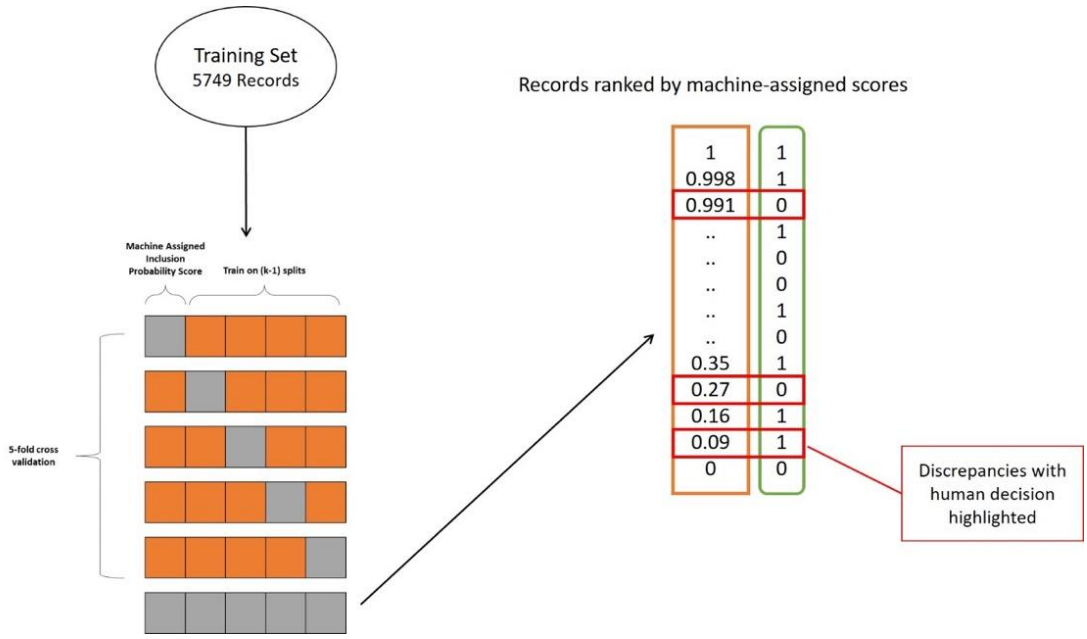


Figure 3.2. Error Analysis. The methodology for using cross-validation to assign ML predicted probability scores. The ML predicted probability scores for the records were checked against the original human inclusion decision.

After the errors in the training set were investigated and corrected as described above a new model was built on the updated training data. The outcome of error analysis is presented as reclassification tables, the area under the curve (AUC) being used to compare the performance of the ML algorithm trained on the 'old' training set of records, and the net reclassification index (NRI) (Kerr et al., 2014) used to compare the performance of the classifier built on the updated training data with the performance of the classifier built on the original training data. The following equation was used:

$$NRI_{\text{binary outcomes}} = (\text{Sensitivity} + \text{Specificity})_{\text{second test}} - (\text{Sensitivity} + \text{Specificity})_{\text{first test}}$$

(Pencina et al., 2008)

The AUC was calculated using the DeLong method in the 'pROC' package in R (Robin, 2017).

Further, I applied the same technique as above to identify human screening errors in the validation dataset. Due to the small number of records in the validation set (1251 records), it was assumed that every error would be likely to impact measured

performance, and so the manual screening of the validation set involved revisiting every record where the human and machine decision were incongruent. The number of reclassified records was noted. The inter-rater reliability of all screening decisions on training set and validation set between Reviewer 1 and Reviewer 2 were analysed using the 'Kappa.test' function in the 'fmsb' package in R (Nakazawa, 2018).

3.3 Results

In this section I first describe the performance from the ML algorithms. I then show the results from the analysis of human error, and finally describe the performance of the ML algorithm after human errors in the training and validation set have been corrected.

3.3.1 Performance of Machine Learning Algorithms

Table 2 shows the performance of the two machine learning approaches from the SLIM (Systematic Living Information Machine) collaboration. The desired sensitivity of 95% (including lower bound 95% CI) has been reach by both approaches. Both approaches reached 98.7% sensitivity based on learning from a training set of 5749 records, with an inclusion prevalence of 13.2% (see below). Approach 1 reached a higher specificity level of 86%. This is visualised on an AUC curve (figure 1).

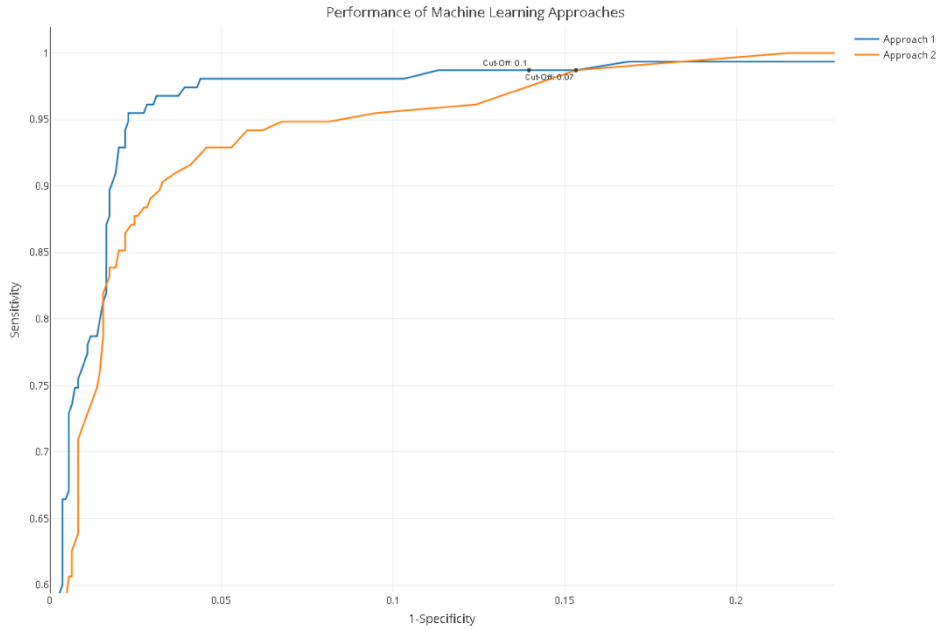


Figure 3.3 Performance of Machine Learning Approaches.

For the interactive version of this plot with cut-off values, see code and data at https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/ML-fig3.html

Table 3.2 Performance of machine learning approaches on depression training dataset.

	Approach 1	**Approach 2**
Training Set Size	**5749**	**5749**
Optimal Cut-Off Score	**0.1**	**0.07**
Sensitivity	**98.7%**	**98.7%**
Upper 95% CI	**99.7%**	**99.7%**
Lower 95% CI	**9.49%**	**94.9%**
Specificity	**86.0%**	**84.7%**
Precision	**50%**	**47.66%**
Accuracy	1096/1251 = **87.6%**	1081/1251= **86.4%**
WSS@95%	**0.705**	**0.693**

3.3.2 Error Analysis & Reclassification

Cohen's κ was run to determine the interrater agreement of screening decisions between Reviewer 1 and Reviewer 2. K = 0.791 (95% CI, 0.769 to 0.811), p < 0.0001, with 281 records requiring a third reviewer decision. To assess whether machine learning algorithms can identify human error and therefore improve the training data, error analysis was conducted. Seventy-five papers out of 5749 papers had predictive scores very far from the human assigned labels, so were reassessed to see if these were due to human errors. Out of 75 rescreened papers, the machine corrected the human decision 47 times. The machine was wrong, (i.e. the initial human decision was correct) 28 times. The validation set was also rescreened. Ten papers out of the 1251 records were identified as potential human errors. Out of 10 errors, the machine corrected 8 human decisions. These 8 records were all falsely excluded by the human and were now included. The initial human decision was correct twice.

To calculate human error in the training set, the number of errors identified (47) out of the training set (5749 records) was calculated to be at least 0.8%. Of the 47 records reclassified, 11 records were falsely included in the original screening process and were now correctly excluded, and 36 records were falsely excluded in the original screening process and were now correctly included. The machine correctly identified human screening errors, which were calculated to be just under 1% of the dual screened training set. Forty-seven papers out of 760 were 'correctly' reclassified, 6% of the included papers.

Similarly, the human error rate in the validation set (1251 records) was 0.6%. Again, looking at the prevalence of inclusion in this dataset (155/1251), which is 12.4%, the 8 records of out the now 163 were correctly reclassified which is 4.9% reclassified. All 8 records we falsely excluded in the original screening process and are now correctly included.

Test 1: 98.7% + 86% = 184.7%

Test 2: 98.2% + 89.3% = 187.5%

NRI = 3.2%

The updated validation set is considered the new gold standard as 8 records were now included. The confusion matrix for the performance of the machine learning algorithm after the error analysis update on the training records is displayed below in table 3.

Table 3.3 Reclassification of records in validation after error analysis

Test 1 – Original Machine Learning Algorithms results				
Test 2 – Post-error analysis ML results		In	Out	Total
	In	153 160	153 116	306 276
	Out	2 3	943 972	945 975
	Total	155 163	1096 1088	1251

Analysing the human errors identified by the machine learning algorithm and correcting for these errors and re-teaching the algorithm leads to improved performance of the algorithm, particularly its sensitivity. This can save considerable human time in the screening stage of a systematic review. Consider the remaining approximately 64,000 papers, if the ML algorithm results are 3% more accurate, that is approximately 2000 papers that are correctly ‘excluded’ that would not be forwarded for data extraction.

3.3.3 Error Analysis: Improving Machine Learning

Using the error analysis technique above, of the 47 errors identified in the full training dataset of 5749 records, 0.8% were corrected. We retrained approach 1 on the corrected training set and measured performance on the corrected validation set of 1251 records as we consider this to be the 'new' gold standard. The performance of the original approach 1 and updated approach 1 was assessed on the corrected validation set of 1251 records. The performance of this retrained algorithm in comparison to the performance of the original classifier 5 on the updated validation set is shown in table 4.

Table 3.4 Performance of machine learning approach after error analysis.

	Updated Approach 1	Original Approach 1
Cut-Off	**0.09**	0.10
Sensitivity	**98.7%**	98.7%
Upper 95% CI of Sensitivity	**99.7%**	99.7%
Lower 95% CI of Sensitivity	**94.9%**	94.9%
Specificity	**88.3%**	86.7%
Precision	**55.9%**	52.61%
Accuracy	**89.7%**	88.2%
WSS@95%	**961/ 1251 – (0.05) = 0.718**	945/1251 – (0.05) = 0.705

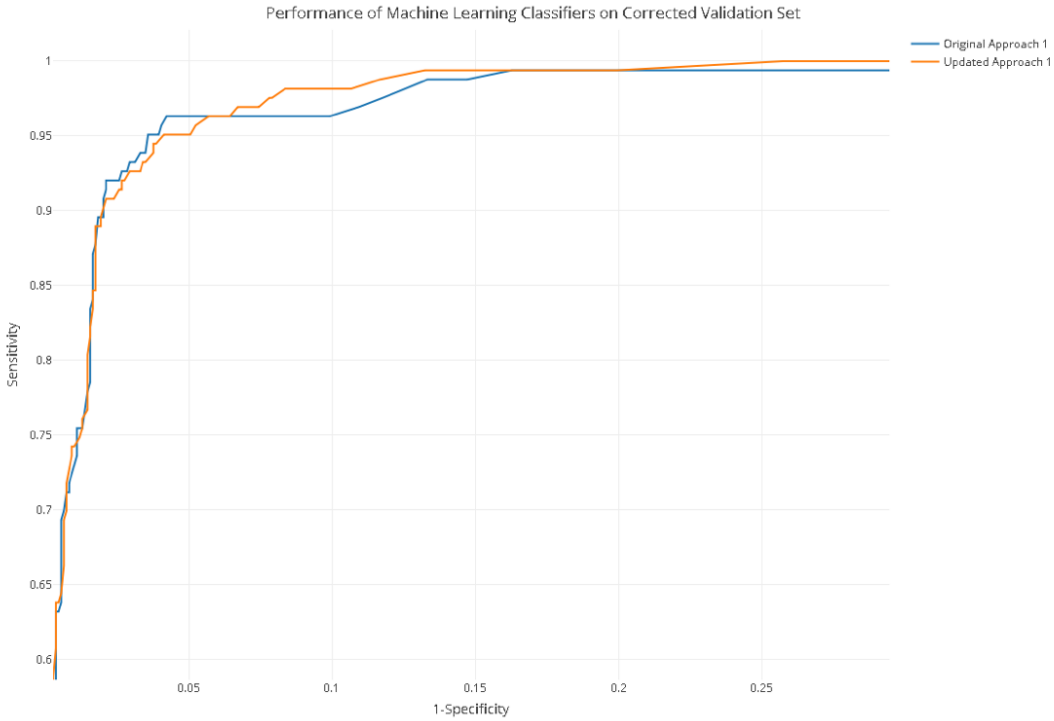


Figure 3.4. Performance of Approach 1 after error analysis. The updated approach is retrained on the corrected training set after error analysis correction. Performance on both the original and the updated approach is measured on the corrected validation set (with error analysis correction). For the interactive version of this plot with exact cut-off values, see code and data at https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews/blob/master/error-analysis-plot.html

I compared the area under the ROC curve for the original approach 1 and the updated approach 1. The AUC for the original approach 1 was 0.9272 (95% CI calculated using DeLong method; 0.914-0.9404). The AUC for the updated approach 1 was 0.9355 (95% CI calculated using DeLong method; 0.9227-0.9483). DeLong's test to compare the AUC between the ROC of the two approaches was applied ', Z = -2.3685, p = 0.0178.

3.4 Discussion

3.4.1 Document Classification

As shown here, machine learning algorithms to have high levels of performance, with 98.7% sensitivity and 88.3% specificity; this sensitivity is comparable to two independent human screeners. The objectives for selecting ML approaches in this project was to achieve a minimum 95% sensitivity (including lower bound confidence intervals), to minimise the number of potentially relevant papers which are wrongly excluded. Thereafter, algorithms were then chosen on the basis of their specificity, to reduce the subsequent human time required to sort through and assess papers.

The two approaches have similar performance. The slight differences may reflect the method of feature generation. These algorithms have high performance on this specific topic of animal models of depression. As demonstrated previously, the performance of various classifiers can alter depending on the topic and specificity of the research question (Howard et al., 2015).

In this study, the cut-off points were selected using the decisions on the validation set to achieve the desired performance. Although this allows the measurement of the maximum possible gain using a given approach in an evaluation setting, in practice (e.g. when updating a review), the true scores would not be available. The problem of choosing a cut-off threshold, equivalent to deciding when to stop when using a model for prioritising relevant documents, remains an open research question in information retrieval. Based on their experience with a given tool, a reviewer may come up with a heuristic fitting their workflow, e.g. if no new includes are seen in the 100 highest-ranked documents, then everything else could be discarded as well. More sophisticated approaches have also been tested (Cormack & Grossman, 2016), but they do not guarantee achieving a desired sensitivity level. It has to be noted that ML-based prioritisation could be useful even if no cut-off is used and all documents are screened manually, since seeing the relevant documents first can help to organise the process and thus reduce the workload (O'Mara-Eves et al., 2015). In a similar broad preclinical research project in neuropathic pain it took 18 person months to screen 33,814 unique records – based on these numbers it would take an estimated 40 person months to screen 70,365 unique records. Performance of machine learning

tools demonstrated here can greatly reduce the amount of human resource needed for initial title and abstract screening of a large corpus of records retrieved from a broad search.

3.4.2 Error Analysis

By using the ML algorithm to classify the likelihood of inclusion for each record in the training set, we highlighted discrepancies between the human inclusion or exclusion decision and the machine decision. Using this technique, we identified human errors, which were then corrected to update the training set.

Human screening of the training set was conducted using the “majority vote” system; it is interesting to consider the potential reasons for errors or ‘misclassifications’ arising in this process. Reviewers’ interpretation of the “breadth” of this wide review might be one contributing factor to discrepancies. With a less clear cut-off, reviewers are unsure of where some articles should be included. Discrepancies arising where Reviewer 1 was more inclusive and where Reviewer 2 was less inclusive, thereby Reviewer 3 will be the deciding factor. A different approach whereby Reviewer 1 and 2 discuss discrepancies might be a pinpoint the exact reasons for misunderstandings or different interpretations of the inclusion criteria. However, for larger projects when using a crowd-sourcing approach with many individual people contributing to each Reviewer, this may not be a practical solution.

We have successfully identified human screening errors which were calculated to be just under 1% of the training set which was dual screened by two independent human reviewers. The prevalence of inclusion in this training set is 13.2% (760 out of the 5749), so an error of 0.8% is likely to be important Therefore errors of false inclusion or exclusion in the training sets may have a substantial impact on the learning of the ML algorithm. This error analysis results in a 3% increase or change in sensitivity and specificity, with increased precision, accuracy, and work saved over sampling of the algorithm. We observed an increase in specificity of 1.6% without compromise to sensitivity. In a systematic review with this number of records this saves considerable

human resources, as the number of records required to screen reduces by at least 1125.

This error analysis was an initial pilot with stopping criteria where if the initial human decision was correct five consecutive times, further records were not reassessed. It is possible and likely that there are further errors in the human screened training set. A more in-depth analysis of the training dataset, investigating every instance where the human and machine decision were incongruent, might identify more errors and further increase the precision and accuracy of machine learning approaches, further reducing human resources required for this stage of systematic review. We have shown here that even with minimal intervention (only assessing incongruent records until the original human decision was correct 5 consecutive times), the performance of ML approaches can be improved.

3.4.3 Limitations & Future Directions

Here we show the best performing algorithms for this dataset with a broad research question. Other dissimilar research questions or topics may require different levels of training data to achieve the same levels of performance or may require different topic modelling approaches or classifiers. The best performing algorithm outlined here, is being applied in an ongoing research project, therefore the 'true' inclusion and exclusion results for the remaining 63,365 records is not yet known. The 'true' results will unfold with the fullness of time.

These machine learning algorithms are deployed in an existing systematic review online platform, EPPI-Reviewer (Thomas et al., 2010), and are in the process of being integrated into the Systematic Review Facility (SyRF) tool, which is focused on the preclinical domain (www.app.syrf.org). This will improve the ease of use of machine learning functions for systematic reviewers, increase the usage of machine learning algorithms for systematic review and significantly reduce the amount of human resources required to conduct systematic review across a range of topics. By allowing a degree of user control over which classifiers and the levels of performance are required for each specific research project. With a broad collaboration such as SLIM

we aim to test many ML algorithms across a range of research topics to identify which classifiers perform best under which circumstances, to be able to provide recommendations to users of SyRF.

This paper outlines a pilot approach to using machine learning algorithms to identify human errors in current systematic review methodology. Future research can investigate this concept more thoroughly by setting up a more comprehensive experimental design. After further investigation into the extent of human error in dual reviewing, the picture will be clearer as to the scale of human error and to what extent a machine learning algorithm can identify and aid in rectifying this. These tools can could be integrated into systematic review platforms, such as SyRF (www.app.syrf.org), and may provide feedback to the systematic reviewer during screening, and could ultimately flag incorrectly screened records as the human screens them for inclusion in a dataset for machine training.

3.4 Conclusions

As shown here, machine learning techniques can be successfully applied to an ongoing, broad pre-clinical systematic review. Machine learning techniques can be used to identify human errors in the training and validation datasets. Updating the learning of the algorithm after error analysis improves performance. This error analysis technique requires further detailed elucidation and validation. These machine learning techniques are in the process of being integrated into existing systematic review applications to enable more wide-spread use. In future, machine learning and error analysis techniques that are optimised for different types of review topics and research questions can be applied seamlessly within the existing methodological framework.

3.5 Availability of Data & Materials

The training and validation datasets, error analysis datasheets, as well as all the records in the depression systematic review are available on Zenodo: DOI 10.5281/zenodo.60269

The protocol for the systematic review of animal models of depression is available from: http://onlinelibrary.wiley.com/doi/10.1002/ebm2.24/pdf

The protocol for the Error Analysis is available via the CAMARADES website and can be accessed directly from this link: https://drive.google.com/file/d/0BxckMffc78BYTm0tUzJJZkc1alk/view

The results of the classification algorithms and the R code used to generate the results is available on GitHub: https://github.com/abannachbrown/The-use-of-text-mining-and-machine-learning-algorithms-in-systematic-reviews.

4 METHODS DEVELOPMENT – AUTOMATION TOOLS TO AID DOCUMENT CATEGORISATION & GROUPING

4.1 Introduction

Broad, shallow systematic reviews are a useful tool to synthesise evidence of a field and provide an overview of a field. Systematic reviews provide an overview of the range of treatments investigated, the different methods used to induce and investigate depressive-like phenotypes. Quantitative synthesis of data from primary studies included in a review, using meta-analysis, can provide insight into the experimental design circumstances under which an intervention is effective in animals. This method can highlight gaps in the literature for further investigation. This methodology has been used extensively by the CAMARADES research team to provide an overview of the literature reporting animal models of stroke, neuropathic pain, schizophrenia, Alzheimer's disease, glioma, breast cancer, spinal cord injury, Parkinson's disease, and multiple sclerosis (Sena et al., 2010; O'Collins et al., 2006, Currie et al., 2013; Seretny et al., 2014; Bahor et al., 2016, Egan et al., 2016; Jue et al., 2018; Chen et al., 2016; Watzlawick et al., 2016; Rooke et al., 2011; Vesterinen et al., 2010). This methodology is being applied to the field of animal models of depression (Bannach-Brown et al., 2016; Kara et al., 2017).

Systematic reviews, while beneficial to the field, require substantial human resources to complete. This gives rise to a number of factors that threaten the usefulness of the findings of a systematic review. Firstly, the longer the review takes the more out of date the results are when published. Usually, larger reviews take longer to complete and therefore the results are published later. Combined with the ever-increasing amount of literature being published in biomedical sciences, this provides a challenge for researchers to get an overview of a field, both for making decisions about experimental design of primary studies, informing clinical trial design, as well as researchers on the boards of funding bodies who are required to be informed of the latest findings in a broad range of topics.

Some of these challenges are addressed by applying automation tools to systematic review methodology. The intersection between biomedical sciences and text-mining and machine learning fields is a fast-evolving field where many new programmes and techniques are being tested. The speed of innovation in this new field, and the collaboration between the two domains is accelerated by sharing tools and data, and the exchange of learning across fields. The application of these automation tools to the domain of preclinical systematic review is a particularly novel study.

Automation techniques can be applied to the screening phase and the annotation phase and are being explored for use in data and outcome extraction. Applying automation tools to annotating documents with key-words can help aid categorisation and grouping of documents. This is particularly useful in broad, shallow systematic reviews where many documents are included in the review and they span across a broad domain. After the screening stage has been completed, the next stage is to prioritise topics for data extraction stage, based on the areas of interest in the field. One approach is to map out and understand the main areas described in the literature, for example, the key PICOs of the studies in the review, the Participants, the Intervention, the Comparison group, and the Outcomes investigated. The PICOs of an animal study can be broadly categorised as the Disease or biological mechanism investigated, any Intervention or drugs investigated, Sample size used in the study, details regarding the Cohort of animals, and the Outcomes investigated (DISCO).

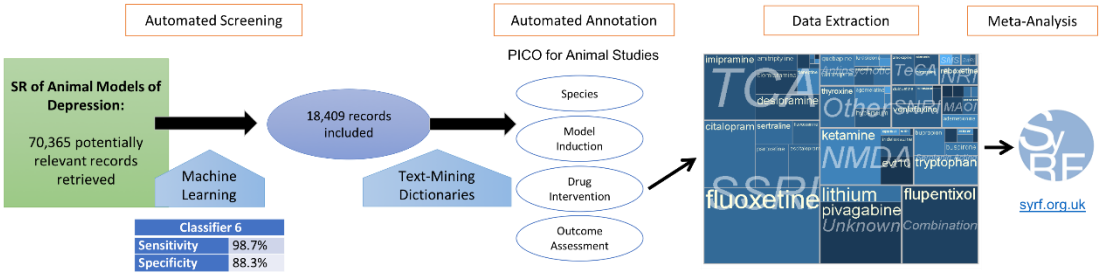


Figure 4.1. Diagram of Workflow and Automated stages of a Systematic Review

Several semi-automated techniques exist to aid the PICO extraction process (Jonnalagadda et al., 2015). These include extracting various factors about the Participants, the Intervention, the Comparison group, and the Outcomes investigated; from more broad aspects such as the disease studied to more detailed information such as sample size and units of measurement. Automation techniques have primarily been applied to texts about clinical trials. These techniques have been shown to be effective in extracting these pieces of information from titles, abstract, title & abstract, as well as full-text articles (Jonnalagadda et al., 2015). A number of approaches have been used to extract these key pieces of information. Some of these approaches have been implemented as tools.

One such technique is the use of single regular expression (regex) terms or regex dictionaries to search the title, abstract, and full text of relevant documents. Regex dictionaries have been applied primarily to aid systematic reviews of clinical studies, to extract PICO information and for assisting in assessing the measures taken to reduce the risk of bias.

One application of regular expression dictionaries for extracting PICO aspects (Patient population & Comparison group) from abstracts of randomised clinical trials has been applied by Hara & Matsumoto (2008). Hara and Matsumoto achieved 90% accuracy in analysing 200 abstracts. A similar approach using regular expressions and the tool 'gazetteer' extracted more detailed information regarding the participants such as the age, gender, and the number of patients from abstracts on cancer (Kelly & Yang, 2013). Kelly and Yang achieved performance, as measured by F-scores, of 87%-100% for the various aspects from the analysis of 386 abstracts from a 'cancer and soy' PubMed search. These tools are applied to the abstracts of articles of interest. Kiritchenko and colleagues used a text-classifier approach to extract information from full-text articles (2010). They used the tool 'ExaCT' (http://rctbank.ucsf.edu/home/exact/exact-user-guide) to extract the information about the dose and frequency of treatment, the patient sample size, eligibility criteria for inclusion in the study, the funders of the study and the primary and secondary outcomes. They achieved a precision of 88% based on 50 full-text papers of randomised controlled trials, with the automation tool trained on 1050 documents (Kiritchenko et al., 2010). Hsu and colleagues (2012) aimed to identify whether a

classifier could classify a sentence as containing key information pertaining to the hypothesis, the statistical methods used such as confidence intervals and significance level, and information about the outcome of the study and conclusions. Using regular expressions to classify sentences, they had a precision of 86% and recall of 78% on 7 papers, with the automation approach trained on 42 full-text papers on chemotherapy for lung cancer. For a full review of all approaches currently being implemented in systematic reviews of randomised controlled trials and synthesis of clinical evidence, see Jonnalagadda et al., 2015.

Similar approaches are applied to extract risk of bias reporting in full-text documents. Millard, Flach & Higgins (2016) tested two different machine learning algorithms to identify whether an article reported allocation concealment, blinded assessment of outcome, and random sequence generation. They used full-text reports of clinical trials in the Cochrane risk-of-bias assessment tool, with manually annotated sentences reporting the risk of bias reporting to train the automation tools. Authors report that algorithms can successfully rank articles by the risk of bias and reduce the human resources required to perform risk of bias assessment in clinical trials.

A slightly different approach to classifying documents has been to understand why documents are classified into a certain category and which sentences in the document are key to understanding this decision. One approach is the use of a technique called convolutional neural network (CNN) which are neural networks with layers that feed forward information and learning to each layer starting with the input and with no hidden layers. This approach allows for more complex data and allows for pooling or weighting of information from each layer (Collobert & Weston, 2008). Zhang, Marshall and Wallace (2016) used convolutional neural networks, to extract the potential sentence with the rationale for the classification of documents. Zhang and colleagues combined the distantly supervised approach of CNN with labelled data, the tagged sentences from user-input were used to improve performance. Their Rationale-Augmented CNN (RA-CNN) has high performance on biomedical risk of bias literature including data from the Cochrane risk of bias dataset (Marshall et al., 2016). A further approach named supervised distant supervision (Wallace et al., 2016b) used semi-structured data from the Cochrane Database of Systematic Reviews which stores PICO information for each clinical trial included in a Cochrane review. Over 12,000

distantly labelled documents and 2,821 sentences directly labelled from 133 papers were used for their novel approach of supervised distant supervised learning (Wallace et al., 2016b). This approach achieved precision levels of between 88.7% and 92.7% on 50 articles, to maximise the information gained from a small amount of human effort.

Automated tools to assess the risk of bias in animal studies have been investigated. Bahor and colleagues use Regex terms to identify the reporting of measures to reduce the risk of bias in experiments of animal models of stroke (Bahor et al., 2017). Bahor and colleagues achieved accuracy levels between 67% and 100% in 964 full-text documents across two different subject areas.

Many of these approaches rely on a dataset of already annotated documents in a specific domain or topic, in order to “learn” the classification. In the approach we outline below, documents are “grouped” by the tags that are labelled as matching the terms of interest. Further, these approaches are primarily applied to the classification of randomised controlled trials of human data, which follow more standardised reporting guidelines in comparison to the reporting of animal studies. Different challenges are present in extracting DISCO or PICO information from animal studies. Studies may investigate several research questions either in the same set of animals, or in separate cohorts of animals, all of which are presented in the same paper. Different studies may use the same outcome assessment tools to measure different behaviours or endpoints.

Further, in the reporting of clinical trials, the publishing culture means that key information is often reported, and therefore it is often possible to discern the PICO from the title and abstract. However, in the literature describing the animal models, not all this information may be discernible from the title and abstract alone. For example, not all outcome measures investigated in the article, or the exact strain of animals used may be mentioned in the abstract, and therefore it is necessary to obtain and assess the full-text version of the article. Retrieving full-text articles is reliant on institutional library subscriptions, bibliographic tools such as EndNote, and human

resources for inter-library loans. In short, the reporting of animal studies varies greatly, and full ascertainment of DISCO information often requires the full-text article.

The aim here was to develop regular expression dictionaries to identify PICO or DISCO information for animal studies. We aimed to automatically ascertain these key pieces of information for grouping and categorising documents, to reduce the human resources for this step. The time reduction of this step allows the next step of the systematic review to be expedited, to select topic areas for further data extraction and meta-analysis.

4.2 Methods

In this systematic review of animal models of depression, we firstly used machine learning classifiers to aid in the screening of 70,365 documents from PubMed and EMBASE. For full methodology of this process, see Chapter 3, Bannach-Brown et al., 2018. After machine-assisted screening, 18,409 documents that were highly likely to be relevant to animal models of depression, a prevalence of 26.16%. The field of animal models of depression is broad; many different techniques are used to model depressive-like phenotypes; different behavioural read-outs are used and the varied investigations of the underlying biological mechanisms behind these phenotypes are carried out. Further, a range of potentially anti-depressant pharmacological interventions are tested in animal models. Therefore, to be able to more efficiently select topics for more in-depth systematic review and potential meta-analysis, categorising these documents into the key areas of interest can help with the next steps of systematic review and potentially, further meta-analysis.

4.2.1 Finding PDFs

The first step was to identify as many full-text PDFs of the documents as possible. The following techniques were used to find PDFs for the 18,409 records. The primary approach was to identify as many PDFs as possible out of the 70,365 documents and match the PDFs up with the included studies. This was due to the PDF retrieval being conducted at the same time as machine learning techniques for screening (see Chapter 3) was being developed and implemented.

Firstly, the ‘Find Full Text’ function in Endnote (v. 7.2.1 (Edinburgh) and v. XT11 (Aarhus)) was used with eproxy links to both the University of Edinburgh and Aarhus University library subscriptions. Eproxy links were added to the Edit > Preferences > Find Full Text. The ‘Find Full Text’ function was run on all 70,365 records. 32,895 full-text PDFs were found.

The open access literature on European PubMed Central (EuPMC) was interrogated using the ContentMine ‘getpapers’ function in node.js (https://github.com/ContentMine/getpapers). The depression PubMed search string was slightly altered for EuPMC (see Appendix 1) and used to interrogate the EuPMC database via their API with the ‘getpapers’ function. 22,003 references were identified. Due to time constraints of the project with ContentMine and memory constraints of the API, the full 22,003 were not downloaded and used in this project.

With many full-text PDFs left to retrieve, a search for further tools to assist the retrieval process was carried out. The search identified a tool called PaperFetch (papertoolbox.com, the website has since been taken offline). This software uses PubMed IDs to search for full-text articles online. Full-texts downloads from the publisher’s websites were enabled with eproxy access. This programme was run on the 62,410 documents. PaperFetch retrieved 19,949 PDFs. Following this, we were contacted by the University of Edinburgh library that access to PubMed and a number of individual journals had been revoked for the IP address the programme was run from, due to a violation of PubMed and journals’ terms and conditions. The programme made requests to PubMed more than 3 times per second. Access to PubMed was revoked for approximately 2 months.

In total, after pairing PDFs with records, these methods retrieved 11,724 full-text PDFs out of the 18,409 included records.

4.2.2 PDF Matching

PDFs and the corresponding records were matched in MS Access using a unique ID generated from the four digits of the year the article was published, and the title. e.g. '2012 - Emotional memory impairments in a genetic rat model of depression: involvement of 5-HT/MEK/Arc signalling in restoration"

4.2.3 Regex Dictionaries

We expanded upon an approach first used in the CAMARADES team to describe measures to reduce the risk of bias in the animal literature of stroke (Bahor et al., 2017). We applied the technique of using regex dictionaries to categorise papers was applied to literature reporting animal models of depression. This literature has initially three key areas of interest that we are interested in mapping: firstly, different techniques used to model depressive-like phenotypes in animals, secondly, different pharmacological therapies used to reduce depressive-like phenotypes in animals, and thirdly, different outcome assessments used to measure depressive-like phenotypes. The first two key areas are discussed here. These dictionaries were built with the help of a fellowship awarded to ABB with the text-mining organisation ContentMine (http://contentmine.org/).

Firstly, commonly used methods to induce depressive-like phenotypes were collated from recently published reviews (Cryan & Mombereau, 2004; Caldarone et al., 2015; Henn & Vollmayr, 2005; Cyran & Slattery, 2007).

Secondly, synonyms and commonly used abbreviations were added. Then, all the known terms were converted to regex format using tutorials in www.regular-expressions.info. Each regex term was tested in www.Regex101.com with text that is known to contain variations of the regex term. The regex terms were adjusted accordingly if there were errors or mismatches. The methods to induce depression in animals were broadly categorised into; genetic inductions, pharmacological inductions such as drug-withdrawal, the use of stress (either in childhood or adulthood) to induce depressive-like phenotypes, and surgical inductions.

The same methodology to create the Regex dictionary for animal models of depression was applied to antidepressants tested in the animals or with animal models. The list of anti-depressants was collected from an open access resource on Wikipedia (https://en.wikipedia.org/wiki/List_of_antidepressants). The links for known trade names of drugs in various countries were followed and these were added to the regex items. The antidepressant drugs were broadly classified into: tricyclic antidepressants (TCAs), serotonin selective reuptake inhibitors (SSRIs), NMDA acting drugs, other serotonin acting drugs, noradrenaline reuptake inhibitors (NRIs), serotonin and noradrenaline reuptake inhibitors (SNRIs), antipsychotics, combination drugs, and drugs acting on other mechanisms and unknown mechanisms.

These regex dictionaries were created in two formats. One that fits with the ContentMine “ami” package, and a second format that fits with the R package created by Jing Liao “shihikoo/AutoAnnotation” (https://github.com/shihikoo/AutoAnnotation/).

The below Table 4.1 presents the format of example entries for the two regular expression dictionaries.

Term	**ContentMine Regex**	**AutoAnnotation Regex**
Citalopram	<regex weight="1.0" fields="citalopram">[cC]italopram|[cC]elexa|[cC]ipramil</regex>	[cC]italopram|[cC]elexa|[cC]ipramil

4.2.4 PDF to Text Conversion

The Github package “AutoAnnotation” developed by Dr Jing Liao was used to convert PDFs to text format. This package uses the R package ‘pdftotext’ to convert readable PDFs into text. This program categorises documents into 4 groups after conversion to text;

Table 4.2 The number of pdfs successfully converted to text.

Conversion Status	**Number of Documents**
Error: Failed to read file	6685
Error: Pdf not found	418
Error: Text file not found	964
OK: File is read Successfully	10342
Total:	18409

The regex dictionaries for model and drug were run through all records, identifying matched regex's in the title, abstract, and where available, in the full-text PDF. The average number of times a regex term occurs in a document and across the whole corpus of documents (18,409 unique records) was calculated using various R packages (v. 3.4, see full code in Appendix 2). Documents where none of the model or drug regex terms occurred were removed. 13,462 of the 18,409 documents had no regex terms from the Model dictionary identified. 15,165 of the 18,409 documents had no regex terms from the Drug dictionary identified.

4.2.5 Application of Regex Dictionaries

The regular expression dictionaries were run on the citations included by machine-assisted citation screening. The number of times a regular expression dictionary term appeared in a document was calculated using the AutoAnnotation R package (Liao, 2017). This package utilises 'gregexpr'the text-mining base package to ascertain the number of times a regular dictionary expression term occurs in a body of text. Two measures were created from this number. 1) The number of documents a term occurs in; 2) The average number of times a regex term occurs in these documents (termed "frequency of occurance"). These two measures were used to create visualiations of the corpus of literature on animal models of depression. Treemap plot visualisations were generated using the 'plotly' (Sivert , 2018; (v. 4.7.1) and 'treemapify' (Wilkins, 2018) packages in R. The category of regex term, for example fluoxetine categorised

as SSRI (see section 4.2.3), we used to group similar drugs or models together in the treemap plot.

4.3 Results

The treemap plots for “Methods of Model Induction” and “Antidepressant Drugs” were generated; see figures 4.2 A and B. These plots display the frequency of regular expression dictionary terms in the animal model of depression corpus. The area of each tile is proportional to the number of documents the term appears in. The colour of each tile is proportional to the average frequency of the term in each document. These plots allow the viewer to ascertain the commonly used models and drugs in the literarute visualliy.

4.3.1 Shiny App Development

This code was developed into a ‘Shiny’ app in R (see app here, https://camarades.shinyapps.io/Preclinical-Models-of-Depression/). The application is interactive, free to use and open access, and connects with the systematic review platform SyRF. Users can select the specific drug or model of interest, view the frequency of the documents, as well as the frequency of terms across the whole corpus, export the studies that have been ‘tagged’ as containing the topic of interest, and import the studies directly into their systematic review project in SyRF to continue their systematic review to completion. Links to the source code are available in Appendix 2.

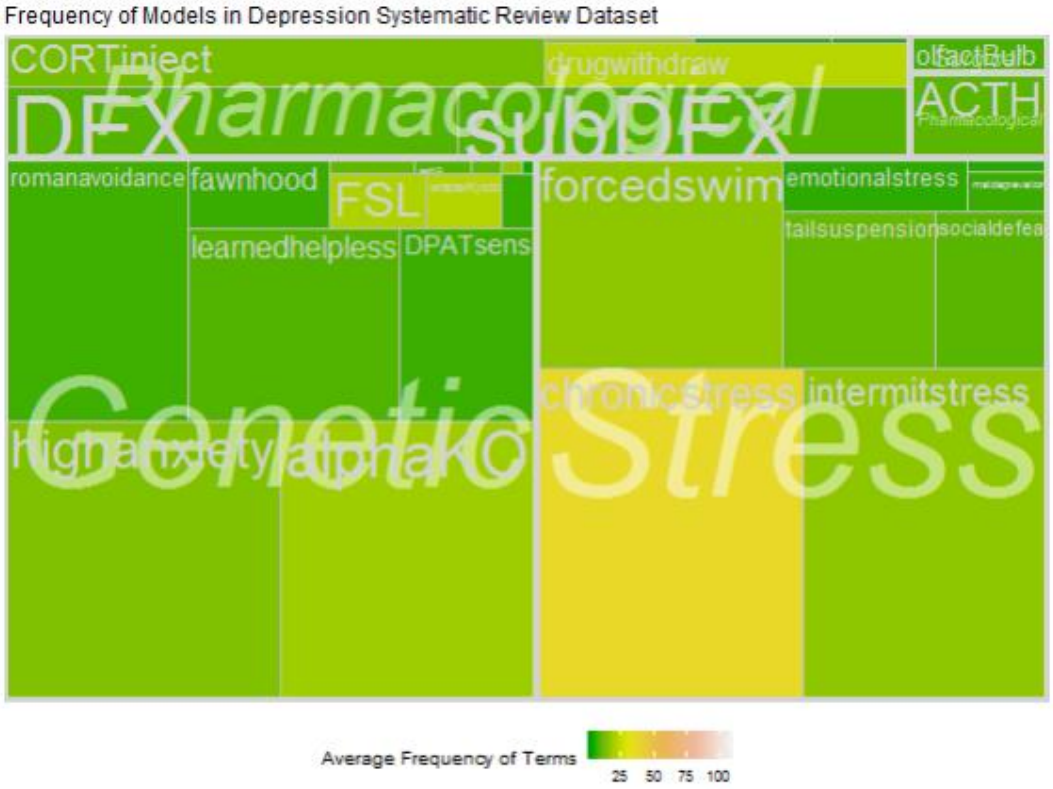


Fig 4.2 A Treemap plot that show the frequency of model terms in the depression corpus. The area of each tile is proportional to the number of documents the term appears in. The colour of each tile is proportional to the average frequency of the term in each document.

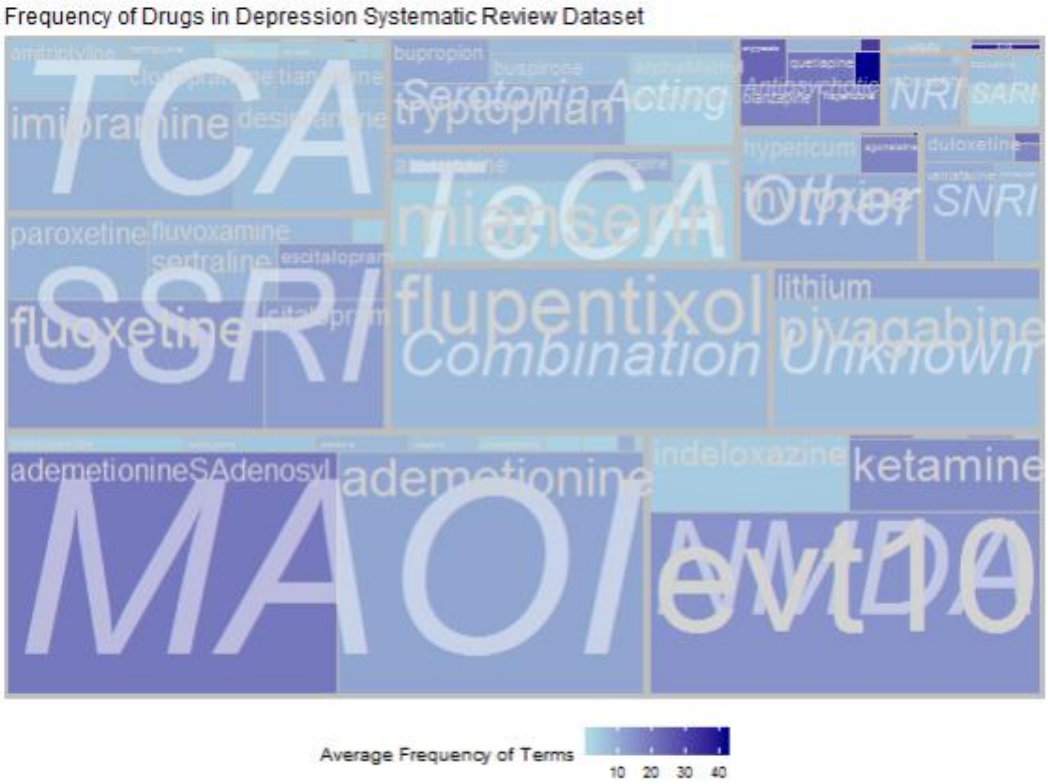


Fig 4.2 B. Treemap plot that show the frequency of drug terms in the depression corpus. The area of each tile is proportional to the number of documents the term appears in. The colour of each tile is proportional to the average frequency of the term in each document.

4.4 Discussion

We have used these regex dictionaries of key DISCO terms for three ongoing reviews, a systematic review and meta-analysis of ketamine as an antidepressant (see preliminary data in Chapter 5), and a systematic review and meta-analysis of the Flinders Sensitive Line model of depression, and a systematic review and meta-analysis of LPS-induced depression. Our approach in these reviews has been to export the list of document IDs where the key term of interest occurred at least once in either title, abstract or full-text PDF, manually check that documents are in fact relevant, before extracting outcome data, experimental design characteristics, and risk of bias information for the studies relevant to the research question. In one step, this tool has greatly reduced time in comparison to manually extracting the topics of interest by hand for all 18,409 documents, before going on to the next step.

This approach is very malleable; any regex dictionary can be used to investigate any range of topics. However, updating these broad dictionaries also requires human input to remain current. Real world concepts and the terminology to describe these concepts change over time, which is what is known as concept drift. Concept drift presents a potential issue for automation tools as it may make the dictionaries more out of date and makes model learning a more complex task as predictions become less accurate with time (Webb et al., 2017). The question is how to capture this drift in key concepts of disease modelling in animals and how to account for this in the use of automation tools to improve preclinical research. In future, the integration of concept drift into our understanding of the reporting of key terms in the literature, through increased collaboration with machine learning and text-mining teams, would help us ensure that dictionaries stay current.

Staying up-to-date with the current literature and novel techniques and approaches within a field is demanding. There is an insurmountable number of papers published on PubMed a year, an average of two papers per minute (Landhuis, 2016). Further, depression is a complex and heterogeneous disease. Understanding the behaviour and underlying biological mechanisms through the use of animal models highlights this. The same outcome measure can be used to assess many different types of behaviour, for example, the open-field test can be used to measure anxiety, locomotion, and exploratory behaviour, and others. The complexity of behaviour, our tools to understand behaviour and biology, the association between the two to identify biomarkers, combined with the lack of standardised terminology or reporting structures captures this issue (Michie & Johnston, 2017). Further, the poor reporting of experimental studies intensifies the complexity (Michie & Johnston, 2017).

One approach to deal with this information overload and the complexity of the research is to improve the currently curated dictionaries and introduce a crowd-sourced effort from many researchers across the field to create a field-wide ontology. With more collaboration with field-specific research groups, we can expand upon the detail of the existing curated dictionaries. A more in-depth understanding of the synonyms used in this field and the various methods of describing the techniques

would enable this approach to continue to be relevant and useful and reduces the waste of research resources simply due to the inadequacy of our quality and structure of reporting of experiments, and the terminology used (Michie & Johnston, 2017). With the aim of making synthesised research more accessible, perhaps improving the search terms for accessing the accumulated evidence or body of evidence that we have on a field, would increase this accessibility (Michie & Johnston, 2017). Creating an ontology, a systematic structure for organising the knowledge, might facilitate investigations into the relationships between these variables, improve our understanding of the field and specific sub-domains, the relationship between sub-domains, and potentially reveal new patterns that emerge through the use of computer-aided investigation (Michie & Johnston, 2017).

The use of ontologies to improve preclinical research is an emerging field. Teams within spinal cord injury have created an ontology for improving the understanding of spinal cord injury which has spurred the generation of new biological hypotheses (Callahan et al., 2016; Callahan et al., 2017). The Federal Interagency Traumatic Brain Injury Research (FITBIR) organisation also has a similar approach to sharing data and creating common dictionaries and ontologies of terminology used. Data sharing and the use of ontologies in spinal cord injury has led to multi-centre animal trials, improved reporting of experimental studies, and the application of novel techniques to answer biological hypotheses (Nielson et al., 2015).

There is a very broad range of both behavioural, physiological and neurochemical read-outs and techniques used in animal models of depression. Collecting and mapping all of these concepts accurately might prove difficult and may only be feasible with a combined effort from researchers in the field. However, experience in other fields demonstrates the utility of these collaborations.

4.4.1 Limitations

A limitation of the current study is the proportion of studies that either did not have a PDF available through university subscriptions or PDFs that were unable to be converted to text because the file was corrupted, or because of the format of the PDF

meant it was unable to be converted to text with the object character recognition software involved in the R packages. In total, the regex dictionaries were unable to be run through the full texts of 8,067 documents. The majority of these documents (6,685 documents) were unable to be found through the university subscriptions. Although university library subscriptions are comprehensive, it is not feasible with publisher subscriptions fees for libraries to have access to all available knowledge, especially when it is behind a paywall. The benefits of open access and data-sharing are discussed in Chapter 2. In future, with a culture shift in academia towards recognising the benefits of open access, this will become less of an issue and all knowledge pertaining to a field will be available to be synthesised.

A second limitation of this study is the extensiveness and breadth of the regular expression dictionaries. There were many documents where none of the terms, from either the model dictionary or the drug dictionary, were found. The reason for this could be that some of the documents may not be relevant to the research question. This is likely as the specificity of the machine learning approach used was 89.7% (see Chapter 3). Further it could be due to not having the full-text PDF available. Where only the title and abstract are available, not all methods to induce depression or all drugs tested may be reported and therefore key information may be missed. In addition, it is likely that the dictionaries are not comprehensive enough, as this is an initial attempt at mapping this field. They will require updating as more resources are invested in mapping the knowledge and extensive dictionaries and ontologies are formed. This process will be expedited with the collaboration of experts in the field to contribute to these ontologies to improve preclinical modelling of depression.

4.4.2 Future Directions

The next steps of this project are to validate the dictionary accuracy by checking the categorisation against a human decision and to investigate if there is a correlation with the number of times a term occurs and how accurate the categorisation is. When using the frequency of each term's occurrence to make predictions regarding the DISCO information, there is potentially an optimal cut-off number for each topic. For example, if a term occurs below 3 or 5 times per full-text document, it might be likely that this drug is not investigated in the study and is merely mentioned in the

introduction or discussion. Another potential approach is to identify whether running dictionaries on sections of an article is more precise than running the dictionaries on the whole article. For example, is the key information regarding model and drug intervention adequately described only in the Methods and Results section of an article, where running the dictionaries only on these sections removes noise?

Future steps for expanding this work is to build a dictionary for the outcome assessments reported in articles describing animal models of depression. Outcome measures of behaviour in animals are complex and often, not all outcome measures assessed are reported in the title and abstract. Therefore, not having access to the full-text PDFs for understanding the breadth of outcome measures used could be a potential challenge to creating an ontology. Here, establishing working groups and crowd-sourced data sharing groups with experts and laboratories in animal models of depression to develop ontologies of our current understanding of the field may be particularly useful.

Although this is an initial attempt at mapping key terms within the field of animal models of depression, mapping the overlap of model induction techniques with the drugs tested in the literature might provide further insights. Investigating which documents contain multiple key-terms, where the terms overlap, may reveal information such as where drugs have not been investigated in certain models, such as genetic models. Highlighting gaps in the literature could lead to the generation of new hypotheses which can be tested.

In future, a link with a repository of preregistered preclinical trials could enable the easy extraction of key DISCO terms. This approach has been used with mapping PICO terms with published clinical trials on PubMed and the original ClinicalTrials.gov trial registration (Kim et al., 2016). They achieved an accuracy of 90% across PICO terms. This is another technique to extract key information that might prove useful to accelerate evidence synthesis.

4.5 Conclusion

This chapter outlines the approach used to create custom regular expression dictionaries of methods to induce depressive-like phenotypes and drug interventions relevant to depression, applying these dictionaries to a corpus of documents, with the aim of categorising and grouping documents into key areas of interest. This approach has already informed three ongoing reviews. This is the first known approach to applying text-mining and dictionary-based approaches to documents of animal models of psychiatry. However, much work is still needed to validate and expand this approach to ensure that the approach is successful in increasing the accessibility of the accumulated evidence, and for use in developing biological hypotheses. I hope that researchers from the field of animal models of depression may contribute to dictionaries and ontologies in future to harness the full potential automation tools to improve preclinical experiments.

5 INTERVENTIONS TARGETING THE GUT MICROBIOTA IN ANIMAL MODELS OF DEPRESSION: A SYSTEMATIC REVIEW AND META-ANALYSIS

This results from this review have been completed with help from Anthony Shek (AS), an honours student at CAMARADES, and Dr. Sarah McCann (SKM), a post-doc at CAMARADES. This chapter reports a systematic review and many of the methods are done in duplicate for good practice. Where steps of the review have been done in duplicate, the contribution will be highlighted by using “we/our”.

5.1 Introduction & Background

Major depressive disorder (MDD) is the leading source of disability globally (Marcus et al., 2012) and treatment-resistance among patients is roughly 50% (Thomas et al., 2013). Therefore, better understanding of the mechanisms behind MDD and the search for potentially effective and novel therapeutic targets are high research and healthcare priorities. Communication between the gut microbiome and the brain may play a role in neuropsychiatric disorders (Cryan & O’Mahony, 2011). The role of gut microbiota in depression is a rapidly growing research field. There has been particular interest in the use of 'psychobiotics' to improve symptoms of low mood in depression. The term 'psychobiotics' is defined as "a live organism that, when ingested in adequate amounts, produces a health benefit in patients suffering from psychiatric illness" (Dinan, Stanton & Cryan, 2013, pg 708). Increasing amounts of research are being conducted to better elucidate the mechanisms under which psychobiotics impact mood and cognition in neuropsychiatric disorders. Current evidence from animal models of depression suggests that psychobiotics communicate using the gut-brain axis and via the central nervous system (CNS), bacterial metabolism and mediating immune signalling.

The gut microbiome primarily communicates bi-directionally with the brain via the vagus nerve, which is the main afferent pathway from the abdomen to the brain (Sherwin et al., 2016). A number of other mechanisms have been explored, as evidence highlights that psychobiotics can impact independently of the vagus nerve, in experiments with vagotomised rodents (Klarer et al., 2014; Bravo et al., 2011).

Mechanisms involved in immune response mediation, tryptophan metabolism, and enteroendocrine signalling have been investigated as potential mechanisms behind depression or the efficacy of treatments for depression. Bacterial commensals have been shown to regulate enteroendocrine signalling, which produce peptides such as serotonin cholecystiokinin (CKK), glucagon-like peptide-1 (GLP-1), and peptide YY (PYY), which are involved in key processes in depression such as limbic responses to fear, neurogenesis, and weight loss/gain (Sherwin et al., 2016). Further, gut microbiota can impact tryptophan metabolism which is involved in serotonin synthesis, a key neurotransmitter implicated in depression and antidepressant responses (Jenkins et al., 2016).

Gut microbiota has been shown to have significant impact on the CNS and neurotransmission involved in depression, including the production GABA which is able to cross the blood brain barrier (Takanaga et al., 2001; Sherwin et al., 2016), dopamine and noradrenaline production in the gut, immunoregulation via histamine, regulating brain derived neurotrophic factor (BDNF) expression, controlling the microglial activation associated with depression, and modulation of the hypothalamic-pituitary-adrenal (HPA) axis activation (Sherwin et al., 2016).

Psychobiotics, in the form of probiotics and prebiotics, are gut microbiota-targeting compounds to reduce depressive-like outcomes in animal models. The definition of probiotics is “a live micro-organism that, when ingested in adequate amounts, produces a health benefit in the host” (Kennedy, Kirk & Gardiner, 2001). Many probiotic strains have been shown to impact behaviour and physiology in human depression and animal models of depression (Abildgaard et al., 2017; Tillmann et al., 2018; Wang et al., 2016). Prebiotics, substrates that are selectively utilised by a host organism providing a health benefit (Gibson et al., 2017), have also been reported to have been investigated in neuropsychiatric outcomes of depressive-like and anxiety-like phenotypes (Burokas et al., 2017; Mika et al., 2017; Thompson et al., 2017; McVey Neufeld et al., 2017). Some of these psychobiotics have been tested in randomised controlled trials (RCTs). Two recent reviews of RCTs identified studies investigating the effects of probiotics (Huang et al., 2016; Romijn & Rucklidge, 2015). Meta-analysis of five studies concluded that probiotics reduced depression symptoms (Huang et al., 2016) and a summary of 10 RCTs concluded that probiotics were

generally more effective than placebo in standardised depression rating scales (Romijn & Rucklidge, 2015). The mechanisms behind the effect of probiotics have in depression are yet to be fully elucidated.

Studies have also investigated the effect of antibiotics, and what impact this has on depressive-like outcomes in animals. Antibiotics are known to impact the composition of the gut microbiome and decrease diversity (Francino, 2015; O'Mahony et al., 2014). It is generally hypothesised that antibiotics negatively impact depressive-like outcomes, via some of the proposed mechanisms above, but further investigations are required to fully understand the implications for mood disorders.

Animal models are beneficial tools to mimic aspects of human depression, to understand the underlying mechanisms behind the disorder and characterise antidepressant interventions. Rodents are a commonly used species to mimic depression, due to the experimental control. Systematic review and meta-analyses of non-human animal data are hypothesis-generating tools that provide an overview of the field and can be used to inform the experimental design. Systematic reviews in other neuroscientific fields, such as Alzheimer's disease, stroke, and multiple sclerosis, have provided evidence as to the components of experimental design that may influence efficacy of an intervention (Egan et al., 2016; O'Collins et al., 2006; Vesterinen et al., 2010).

In this systematic review and meta-analysis we set out to review the efficacy of interventions targeting the gut microbiota in animal models of depression. We investigated interventions that either were aimed at inducing depressive-like behaviour or interventions to reduce depressive-like behaviour. We aimed to provide an overview of the literature. We aimed to describe the quality of the data available, and if possible, meta-analysis to determine study design characteristics that influence efficacy (particular bacterial strains used in psychobiotics, route of administration of psychobiotics), and areas where uncertainty remain and further animal experiments may be provide useful evidence.

5.2 Methods & Materials

The methodology for this systematic review was laid out in a pre-specified protocol published on the CAMARADES repository of Systematic Review Protocols (http://www.dcn.ed.ac.uk/camarades/research.html#protocols) on 13th February 2017.

5.2.1 Search Strategy

Studies of interventions targeting the gut microbiota in animal models of depression were identified from an existing database of studies of animal models of depression. This database was collated using the search string in the protocol outlined in chapter 2 (Bannach-Brown et al., 2016). The original search was carried out in May 2016. This database was searched using the key terms; “microbiota”, “gut microbiome”, “germ-free”, “gut-brain axis”, “probiotic”, “antibiotic”. Abstract screening was carried out by AS and ABB. Any screening discrepancies were resolved by a discussion between the two reviewers. Full-text screening was carried out simultaneously with data extraction by AS and ABB.

5.2.2 Inclusion & Exclusion Criteria

Publications were included if they tested an intervention that claimed to impact or work via the gut microbiome. Interventions could either act as a method of model induction in depression or as a potential treatment of depression, in an *in vivo* rodent experiment. Publications were included where any quantitative outcome of depression had been assessed. Studies were included if there was an appropriate control group and where the sample size, the mean and the variance in each group had been reported for the primary outcome of behaviour. There were no exclusion criteria based on age, sex, weight or method of model induction. There were no exclusion criteria based on dosage, timing or frequency of administration of the intervention. Studies were excluded if only anxiety-like behaviour was investigated, or if genomic, proteomic, metabolic or metabolomics outcomes were the sole outcome, without a behavioural outcome. There were no exclusion criteria pertaining to the language or date of publication.

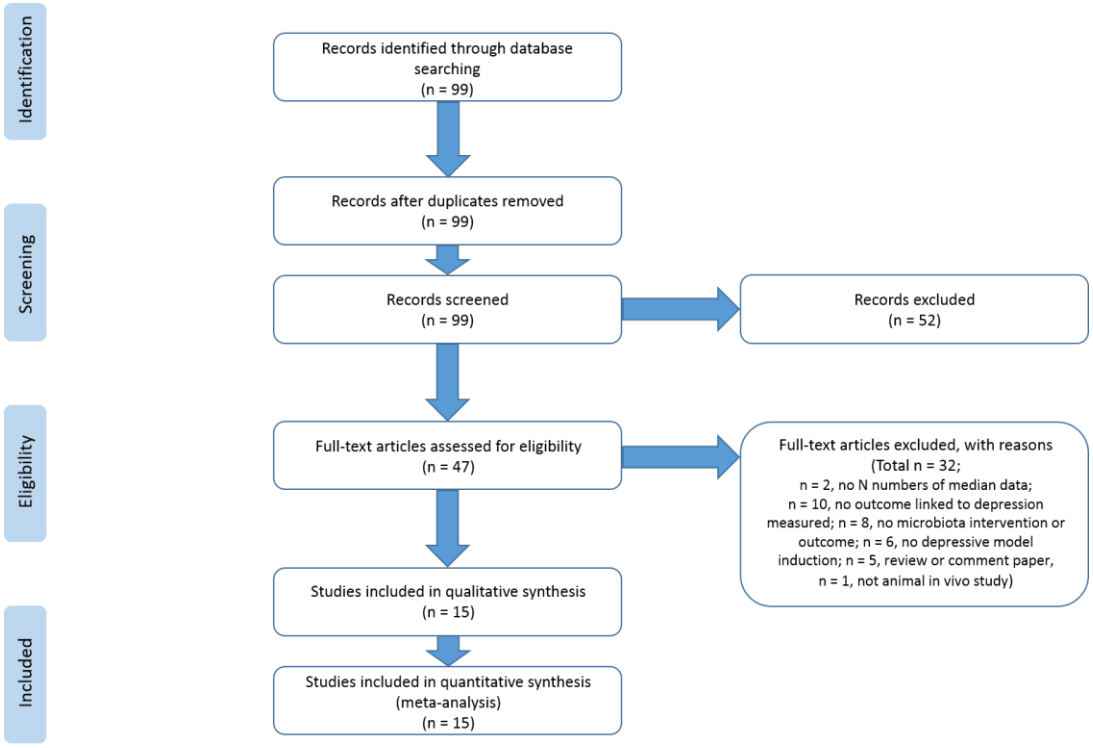


Figure 5.1. PRISMA flowchart of studies included in this review.

5.2.3 Data Extraction from Primary Studies

For each publication, information on quality (see the following section) and experimental design characteristics (animal species, strain and intervention tested) were entered into the CAMARADES Data Manager, a centralised Microsoft Access Database. For all included studies, details regarding the following information was extracted:

- The method of model induction
- The strain, sex and age or weight of animals
- Animal husbandry conditions in particular any germ-free or pathogen free housing conditions
- Information regarding the diet of the animals
- Intervention characteristics such as the dose, the route of administration, the timing and number of administrations where reported

The number of animals, mean and measure of variance (either SD or SEM) for control group and treatment group were extracted for each outcome of interest. The primary outcome measure was efficacy on behavioural scores. Further, data for genetic, microbial, neurochemical, hormonal, anatomical, metabolic, and immunological outcomes were also extracted. See Table 5.1 for classification of outcomes. Where data were presented graphically, universal desktop ruler (https://avpsoft.com/products/udruler/) or the in-built measuring tool in Adobe was used to extract numerical values. Data extraction was carried out in duplicate by AS & ABB, with a third screener (SKM) involved in the reconciliation stage.

Table 5.1 Classification of Outcomes

Outcome Category	**Definition/explanation**	**Example**
Behavioural	Any test in live animals looking at behaviour	Forced Swim Test, Elevated Plus Maze
Microbial	Any measure of bacterial colonies as measured by, for example, mRNA or protein expression	Beta diversity, Alpha diversity in gut microbial colonies
Neurochemical	Neurochemical readouts including neurotransmitter levels or metabolites, neurogenesis, neuronal excitability, as well as brain adrenaline or noradrenaline	BNDF levels, relative mRNA levels
Hormonal	Any measure of hormonal levels, often from plasma	Cortisol, blood adrenaline and noradrenaline levels, testosterone, oestrogen

Anatomical	Any measure of body structure or composition	Body weight, brain structure or size, colon permeability
Metabolic	Any readouts of body metabolism	Glucose or insulin levels or turnover, fecal boli, urine reactivity
Immune	Any physiological responses to foreign bodies (e.g. viruses, fungi, bacteria in the body) as detected via antigens	Cytokines, LPS, T-cell response, Interferon, IL-1, IL-4
Inflammatory	Any physiological response to damage such as trauma, heat, stress, toxins, bacteria – including mRNA or protein expression of these responses	Histamine, Prostaglandin, Bradykinin, Phagocytes
Oxidative Stress	Any oxidative stress response that signals a disturbance between antioxidant defences and production of reactive oxygen species	Antioxidants such as superoxide dismutase, catalase, glutathione. Reactive oxygen species, lipid peroxidation.

5.2.4 Assessment of Quality

Methodological quality was assessed by recording the reporting on any of the following four items; whether a study reported the randomisation of animals to control and treatment groups or model and control groups, whether a study reports the blinding of investigators under the assessment of outcome measures, allocation concealment, and if authors report a sample size calculation (Macleod et al., 2004). Further, we assessed whether authors report that experiments comply with animal welfare regulations, whether authors have included a statement of potential conflicts of interest, and what the source of funding was, if reported.

5.2.5 Data Reconciliation

Two independent reviewers (AS & ABB) extracted data from publications. Extracted data at the level of the publication and the outcome were compared and any discrepancies were checked and corrected. Effect sizes for each comparison were calculated for each reviewers' extracted data and compared. Where effect sizes differed by more than 10%, the original article was checked, and graphs were re-extracted. Where effect sizes for comparisons differed by less than 10%, the mean of the effect sizes and the errors (SEM/SD) were calculated.

5.2.6 Data & Meta-Analysis

For all outcome measures, the effect size measure standardised mean difference (SMD) was calculated for each comparison. A comparison is defined as a pair of cohorts of animals, one cohort that receives an intervention, and the other cohort receives an appropriate control for that intervention (Vesterinent et al., 2014). SMD was used as we cannot infer the 'normal' behaviour or 'normal' biological readouts of an animal, and data are presented on different scales. All equations used are taken from Vesterinen and colleagues (2014). SMD is calculated using Hedge's G standardised effect, as it has a correction factor for small sample size (fewer than 10 animals per group), with the following equation:

$$SMD_i = \frac{(\overline{X}_{c} - \overline{X}_{Rx})}{S_{pooled}} \times \left(1 - \frac{3}{4N-9}\right) \times direction$$

Where 'direction' refers to the correction factor applied to the effect size to account for the direction, whether a higher score in the outcome represents a worse or better outcome (Vesterinen et al., 2014). S_{pooled} is calculated with the following equation:

$$S_{pooled} = \sqrt{\frac{(n_c'-1)SD_c^2 + (n_{Rx}-1)SD_{Rx}^2}{N-2}}$$

Standard error for SMD is calculated using:

$$SE_i = \sqrt{\frac{N}{n_{Rx} \times n_C'} + \frac{ES_i^2}{2(N-3.94)}}$$

Comparisons were combined using random-effects modelling, with Hartung-Knapp adjustment, with a restricted maximum likelihood estimate of between study variance. Random effects meta-analysis assumes that true effects differ between studies and allow us to explore the differences between the studies that contribute to differences in effect sizes. Firstly, the weight of each study is calculated by the inverse of the sum of within study variance and tau-squared (τ^2) as a measure of between study variance:

$$W^*_{+\tau^2} = \frac{1}{(SE^2_{\theta i}+\tau^2)} \quad \text{and} \quad \tau^2 = \frac{Q-df}{C}$$

Q = sum of the squared differences in effect sizes between studies and the pooled effect size ($Q = \sum_{i=1}^{k} W^* \times (ES_{\theta i} - ES_{fixed})2$), df = degrees of freedom, and C = measure used to convert the heterogeneity value into an average and put the value back into original units. The weighted effect size for each study is:

$$ES^*_{rand} = ES_{\theta i} \times W^*_{+\tau^2}$$

The effect sizes from studies are summarised and divided by the sum of weights alone, to get a pooled effect size. The summary effect estimate and the standard error of the estimate is calculated with the following equations:

$$ES_{random} = \frac{\sum_{i=1}^{k} ES^*_{rand}}{\sum_{i=1}^{k} W^*_{+\tau^2}} \quad \textit{and} \quad SE_{random} = \frac{1}{\sqrt{\sum_{i=1}^{k} W^*_{+\tau^2}}}$$

The 95% confidence intervals for the summary effect size are calculated with the following equation:

$$95\%\, CI = ES_{random} \pm 1.95996 \times SE_{random}$$

If several outcomes in the same category, e.g. several measures in the forced-swim test, were reported from the same cohort of animals, these were combined or nested using fixed effects meta-analysis and this aggregate estimate was used for further analysis in the random-effects model. Fixed effects meta-analysis is similar to random

effect meta-analysis, but the assumption is that true effects do not differ between comparisons and that all studies share one true effect size. The fixed effect model assumes that any variation between studies is due to sampling error. The weight of each study is the inverse of the variance. The effect size for each comparison is calculated by multiplying the effect size by the weight.

Weight: $W_i = \frac{1}{SE_i^2}$ Weighted effect size: $W_i ES_i = ES_i \times \frac{1}{SE_i^2}$

The weighted effect size for each comparison is summarised and divided by the sum of weights alone, to give the pooled effect size:

$$ES_{fixed} = \frac{\sum_{i=1}^{k} W^* ES_{\theta i}}{\sum_{i=1}^{k} W^*}$$

The standard error and 95% confidence intervals of the summary effect are calculated with the below equations:

$$SE_{fixed} = \frac{1}{\sqrt{\sum_{i=1}^{k} W^*}} \quad \textit{and} \quad 95\%\, CI = ES_{fixed} \pm 1.95996 \times SE_{fixed}$$

Where multiple intervention groups were used with a single control/vehicle group, the number of animals in the control group was divided by the number of intervention groups. The impact of study design characteristics and study quality were assessed with stratified meta-analysis. These were assessed separately in studies investigating microbiota interventions to induce depression, and studies investigating microbiota interventions as treatments. The study design characteristics pre-specified are outlined in Table 5.2. The models used were grouped into models that directly targeted the gut microbiota (Direct) and models that indirectly targeted the gut microbiota or stipulated that stress impacted the gut microbiota (Indirect). Treatments were grouped into interventions that had proposed antibiotic properties and interventions that had proposed probiotic properties. The underlined outcomes in Table 5.2 were used in the analysis. The reason for not investigating all the pre-specified study design characteristics was that there were not enough comparisons (e.g. only 1 comparison used female animals). The age of animals was not presented uniformly across primary studies. Some studies reported the weight of an animal, whereas other studies reported the age in weeks. These cannot reliably be combined into a single scale. The weight varies greatly between strains and is also different for

animal from different breeders. Therefore, we did not combine this variable. Further, treatment dose was not able to be combined onto a single meaningful scale as some studies administered antibiotics (in mg/kg) and other studies administered probiotics (in Colony Forming Units, CFU).

Table 5.2 Stuidy Design Charcteristics for Model and Treatment Interventions

Study Characteristics for Model Induction Interventions:	**Study Characteristics for Treatment Interventions:**
Direct or Indirect Gastrointestinal Manipulation	Type of Treatment Intervention (Probiotic or Antibiotic)
Strain of Rodent	Treatment Dose
Age of Animals	Route of Administration
Sex of Animals	Number of Times Administered
Number of times the outcome was assessed	Treatment given pre or post model induction

Univariate meta-regression was used to investigate treatment dose as a possible source of heterogeneity. Meta-regression and stratified meta-analysis was performed using a Meta-analysis Online Platform based on R (code available here: https://github.com/qianyingw/meta-analysis-app).

Stratified meta-analysis partitions heterogeneity within groups of similar studies, and between groups of studies. Random effects effect size and Q are calculated for each group and grouped into similar studies and subtracted from the total heterogeneity; residual heterogeneity between groups remains. The heterogeneity between groups is tested using the chi-square test. Meta-regression is a weighted linear regression describing the line of best fit between the effect size and the covariates added to the model. Meta-regression takes into account the within-study variance and the between study variance (τ^2). For meta-regression, τ^2 is calculated using the restricted maximum likelihood estimate (Thompson & Sharp, 1999). Adjusted R^2 is the variance

in the outcome measure that is accounted for by the independent variable. Changes in τ^2 reflect the variance that is explained by the covariates and therefore residual heterogeneity. An F-ratio is calculated to ascertain how much the addition of covariates has improved the prediction of the model.

A simulation study by Wang et al., (2018) at CAMARADES identified that stratified meta-analysis using SMD has low statistical power to detect the effect of a variable of interest, but a low false positive rate, so we can be confident in any significant results found. A Holm-Bonferroni correction was applied to correct for multiple testing. The significance level for the analysis of study design characteristics in modelling experiments was adjusted to $p < 0.025$. The significance level for the analysis of study design in treatment experiments was adjusted to $p < 0.010$. The significance level for the analysis of measures to reduce the risk of bias in modelling and treatment experiments was adjusted to $p < 0.017$.

All effect size measures are reported as standard difference (SD) with upper and lower 95% confidence intervals.

5.2.7 Publication Bias

The risk of publication bias was assessed using visual inspection of a funnel plot, Egger's regression (Egger et al., 2007) and trim and fill analyses (Duval & Tweedie, 2000) were used to identify potentially missing studies.

5.3 Results

5.3.1 Identifying Publications

A systematic search was conducted in PubMed and EMBASE in May, 2016. 70,365 unique publications were identified from the search. Machine-learning algorithms were employed to screen the studies based on a sub-set of documents with human decisions (2 independent reviewers with 3rd for reconciliation). A performance of 98.7% sensitivity and 88.7% specificity was achieved. 18,409 documents were included in the review by the machine-assisted approach, which formed the depression database.

99 studies were identified in the database as containing the microbiota keywords of interest. 47 publications were included at title and abstract screening. 15 publications were included in this review after full-text screening.

5.3.2 Microbiota Interventions as treatments in animal models of depression

10 studies reported a microbiota-targeting intervention to rescue depressive-like behaviour where a model of depressive-like behaviour had been induced. These interventions to treat depressive-like behaviour were grouped into interventions that had proposed antibiotic properties and interventions that had proposed probiotic properties.

5.3.2.1 Intervention Variables

Of interventions that had probiotic properties, no two products or strains were the same. Products or strains of bacteria used were (8 different probiotic treatments): Probio'Stick, L. salivarius HA113, L. rhamnosus (JB-1), Bifidobacterium infantis 35624, L. helveticus NS8, L. plantarum PS128, B. Longum 1714, B. Breve 1205, and fecal samples from healthy control patients given once (fecal colonisation). Of interventions that had antibiotic properties (3 different antibiotic treatments), products used were; minocycline, Streptomycin Sulphate & Penicillin G, and an antibiotic cocktail made up of vancomycin, neomycin, metronidazole, and amphotericin B. These interventions were administered to animals between 14 days and 45 days, once daily, either via oral gavage or orally.

We pre-specified that at least 25 independent comparisons per outcome measure were required for quantitative analysis with meta-analysis. As above, a comparison is defined as a pair of cohorts of animals, one cohort that receives an intervention, and the other cohort receives an appropriate control for that intervention (Vesterinent et al., 2014). There were only a small number of primary studies included in this review, but as this is an emerging field, and we were looking to inform a primary animal experiment, we decided to explore heterogeneity with meta-analysis of the primary outcome, behaviour.

In the 17 comparisons from 9 studies that investigated microbiota-targeting interventions as treatments in behavioural outcomes (621 animals), overall interventions led to a significant improvement in depressive-like behaviour (0.702 SD; 95% CIs [0.3928; 1.0120]), figure 5.2. There was moderate heterogeneity between the studies τ^2= 0.226, I^2 = 68.9%, Q = 51.50, df = 16, p < 0.0001. Sample sizes of model induction groups ranged from 7 animals per group, to 69 animals per group, with a median of 10 animals per group. The impact of study design characteristics on heterogeneity was investigated.

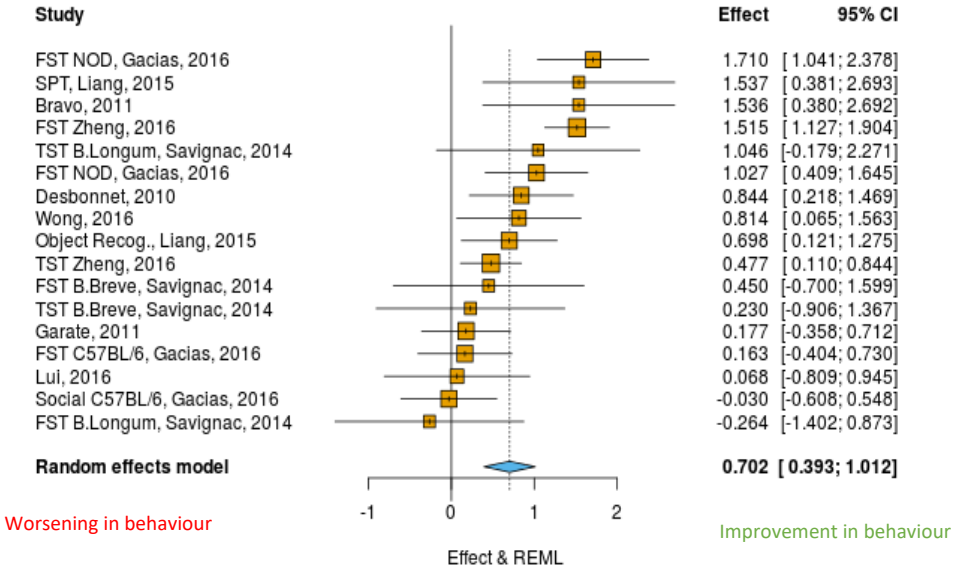


Figure 5.2. Forest plot of the effects of microbiota interventions to reduce depressive-like behaviour in the meta-analysis. Positive effect sizes indicate an improvement in behaviour and negative effect sizes indicate a worsening in behaviour associated with depression. Solid orange squares represent the effect size of each comparison with 95% confidence intervals. The blue diamond represent the pooled effect size, with the edges of the diamond representing the 95% confidence intervals. The solid line represents the line of no effect, if 95% confidence intervals cross this line, the study is not effective. The dashed line represents the global effect.

5.3.2.2 Impact of Study Design

None of the four study design characteristics (drug category, route of administration, timing of administration, or number of administrations) investigated significantly explained heterogeneity between the studies. The study design characteristics for each study are summarised in Appendix 3, Table 1

5.3.2.3 Impact of Measures to Reduce the Risk of Bias

We investigated 4 measures to reduce the risk of bias; random allocation to group, blinded assessment of outcome, allocation concealment, and reporting of a sample size calculation. Random allocation to treatment or control group accounted for a significant proportion of the heterogeneity between studies (Fig 5.3), difference in Q = 7.01, $\chi 2 = 3.84$, df = 1, p = 0.008. Reporting of random allocation of animals to group was associated with greater estimates of effect. Blinded outcome assessment did not significantly account for heterogeneity between the studies. There were not enough studies that reported a sample size calculation (1 study reported sample size calculation) or allocation concealment (no studies reported allocation concealment) to explore these variables as potential sources of heterogeneity. All studies reported compliance with animal welfare regulations and all but one paper reported a conflict of interest statement so these variables were not investigated as potential sources of heterogeneity. Source of funding did not significantly explain differences in effect sizes between the studies. Two studies reported private funding, six studies reported public funding, and two studies reported both public and private funding. An overview of the reporting for measures to reduce the risk of bias for each study is presented in Appendix 3, Table 3.

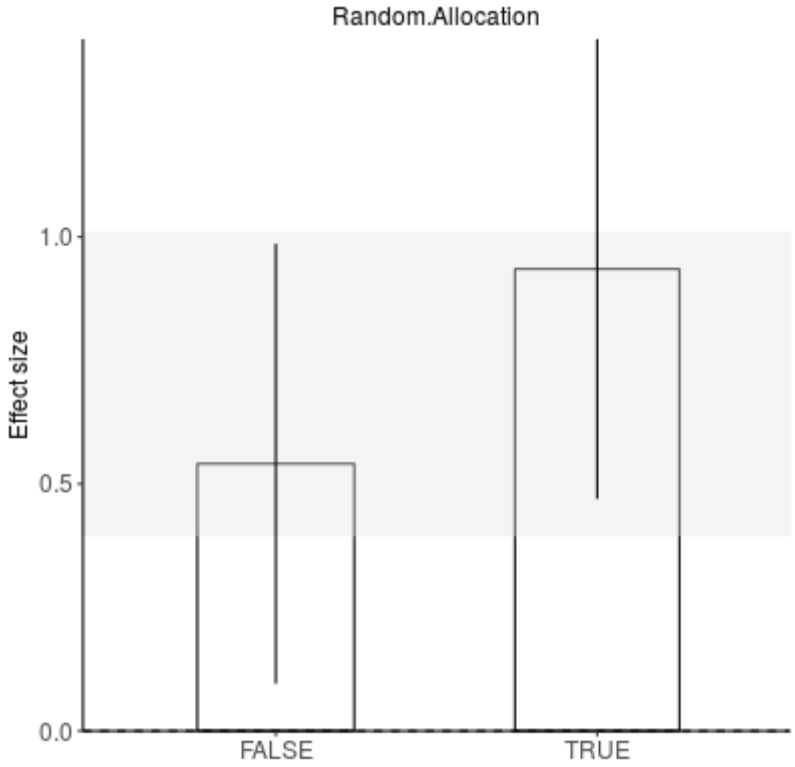


Fig 5.3. Effect sizes associated with random allocation to treatment or control group in experiments. The grey section signifies the overall effect size including 95% confidence intervals. The bars signify the effect size in each sub-group. Errors bars are 95% confidence intervals. 11 comparisons did not report random allocation. 6 comparisons reported random allocation.

5.3.2.4 Other Outcomes

We extracted eight other outcomes from the primary articles; microbial (6 comparisons), neurochemical (67 comparisons), hormonal (16 comparisons), anatomical (8 comparisons), metabolic (3 comparisons), immune (13 comparisons), inflammatory (8 comparisons), and oxidative stress (1 comparisons). Due to the small number of studies, these were not able to be investigated with quantitative analysis. An overview of the outcomes investigated in primary articles reporting gut microbiota-targeting Interventions to reduce depression is presented in Figure 5.4.

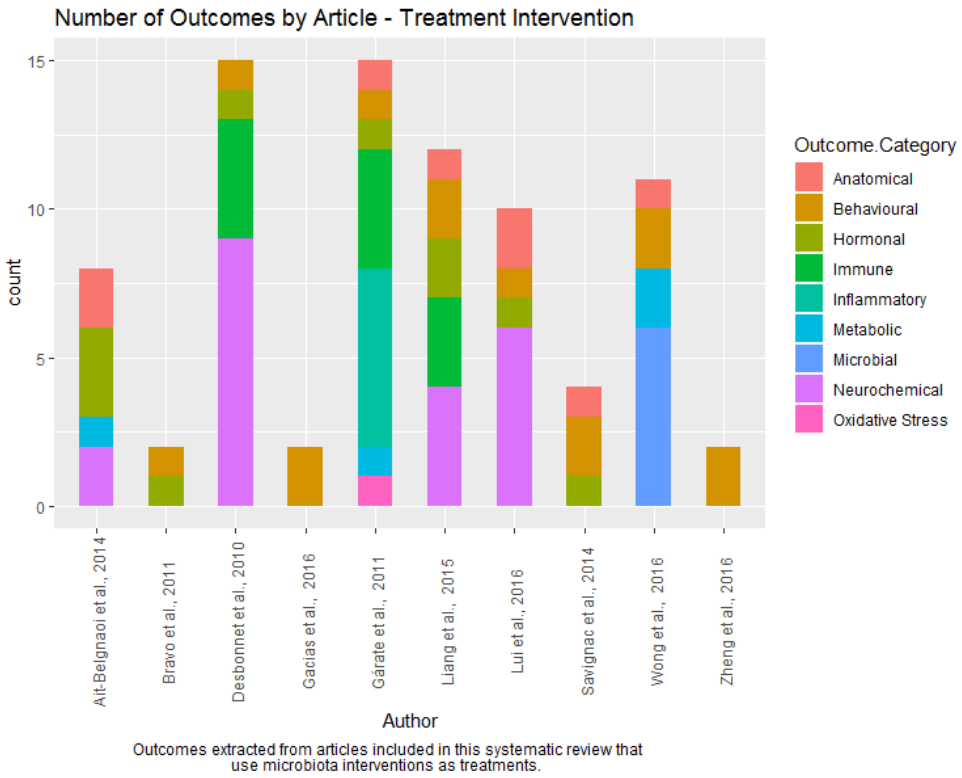


Figure 5.4. The category of outcome measures in experiments investigating gut-microbiota interventions as treatments. Behavioural (17 comparisons), microbial (6 comparisons), neurochemical (67 comparisons), hormonal (16 comparisons), anatomical (8 comparisons), metabolic (3 comparisons), immune (13 comparisons), inflammatory (8 comparisons), and oxidative stress (1 comparisons).

5.3.2.5 Publication Bias

For studies that used gut-microbiota interventions as treatments, there were seventeen comparisons (0.702 SD; 95% CIs [0.3928; 1.0120]). The funnel plot revealed slight asymmetry which was not significant with Egger's Regression.

5.3.3 Interventions targeting the gut microbiota to induce depression

Thirteen studies reported an intervention to induce depression that was stipulated to act through gut microbiome alterations.

Of the 13 studies, 8 studies used mice and 5 studies used rats. All but two studies used male animals, one study used female animals and one study did not report the sex of animals used. The models used were grouped into models that directly targeted the gut microbiota (8 experiments) and models that indirectly targeted the gut microbiota or stipulated that stress impacted the gut microbiota (7 experiments). 13 different models were used across the papers. The direct manipulations of the gut microbiota to induce depressive-like behaviour were; Germ-free housing, magnesium deficient diet, high sucrose diet, high fat diet, ciprofloxacin, gastric gavage, and prolonged weaning. Indirect manipulations of the gut microbiota to induce depressive-like behaviour investigated were; water avoidance stress, forced swim test, maternal separation, chronic mild stress, and chronic restraint stress.

We pre-specified that at least 25 independent comparisons per outcome measure were required for quantitative analysis with meta-analysis. There were only a small number of primary studies included in this review, but as this is an emerging field, and were looking to inform a primary animal experiment, we decided to explore heterogeneity with meta-analysis of the primary outcome, behaviour.

In the 16 comparisons from 9 studies that report microbiota-targeting interventions to induce depression and investigated behavioural outcomes (407 animals), overall interventions led to a significant worsening of depressive-like behaviour (-0.549 SD; 95% CIs [-1.079; -0.019]), figure 5.5. There was high heterogeneity between the studies $\tau^2 = 0.78$, $I^2 = 82.4\%$, $Q = 85.11$, df = 15, $p < 0.0001$. Sample sizes of model induction groups ranged from 6 animals per group, to 44 animals per group, with a median of 14 animals per group. The impact of study design characteristics on heterogeneity was investigated.

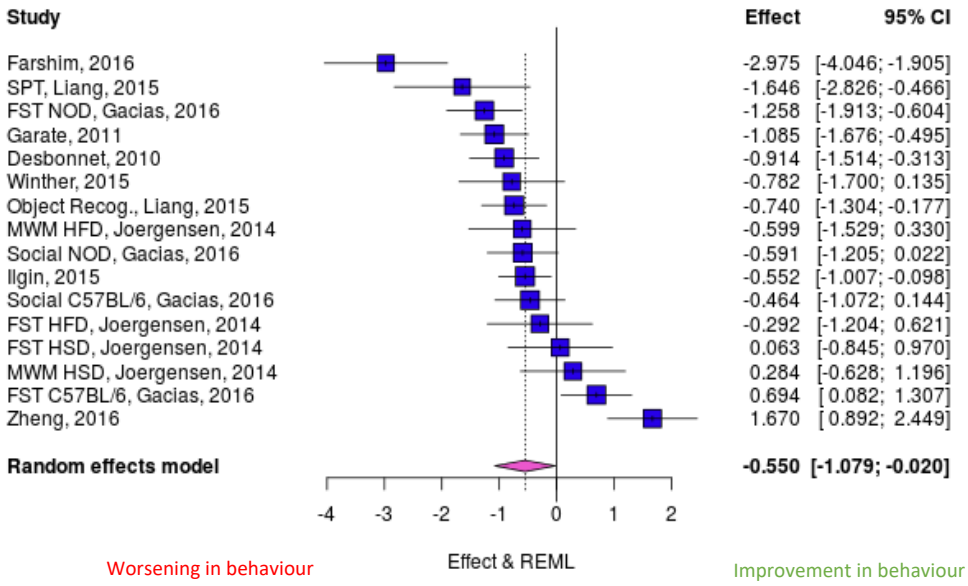


Figure 5.5. Forest plot of the effects of microbiota interventions to induce depressive-like behaviour that was stipulated to act through gut microbiome alterations in the meta-analysis. Negative effect sizes indicate a worsening in behaviour, positive effect sizes indicate improvement in behaviour associated with depression. Solid blue squares represent the effect size of each comparison with 95% confidence intervals. The pink diamond represent the pooled effect size, with the edges of the diamond representing the 95% confidence intervals. The solid line represents the line of no effect, if 95% confidence intervals cross this line, the study is not effective. The dashed line represents the global effect.

5.3.3.1 Impact of Study Design

Three study design characteristics were investigated as potential sources of heterogeneity; model category (direct or indirect), type of animal (rat or mouse), and sex of the animals used. The category of model significantly explained heterogeneity between the studies (difference in Q = 10.23, $\chi 2 = 3.84$, df = 1, p = 0.0013), Figure 5.6A. Studies using indirect models (4 comparisons), the more commonly used models in the field e.g. chronic mild stress, were associated with greater effect sizes (-0.96 SD, [-1.40; -0.52]), whereas direct models (12 comparisons) were not effective at inducing depressive-like effects (-0.38 SD, [-1.07; 0.32]). The type of animal used

in the experiment significantly explained differences in effect size between the studies (difference in Q = 20.11, $\chi 2 = 3.84$, df = 1, $p = 7.31 \times 10^{-6}$), Figure 5.6B. Studies that used rats (6 comparisons) were associated with greater effect sizes (-1.21 SD, [-2.08; -0.33]), whereas no effects were seen in studies using mice (10 studies, -0.129 SD, [-0.739; 0.48]). All but one study used male animals, therefore this variable was not able to be investigated as a source of heterogeneity. A summary of study design characteristics in studies exploring microbiota interventions to induce depression is presented in Appendix 3, Table 2.

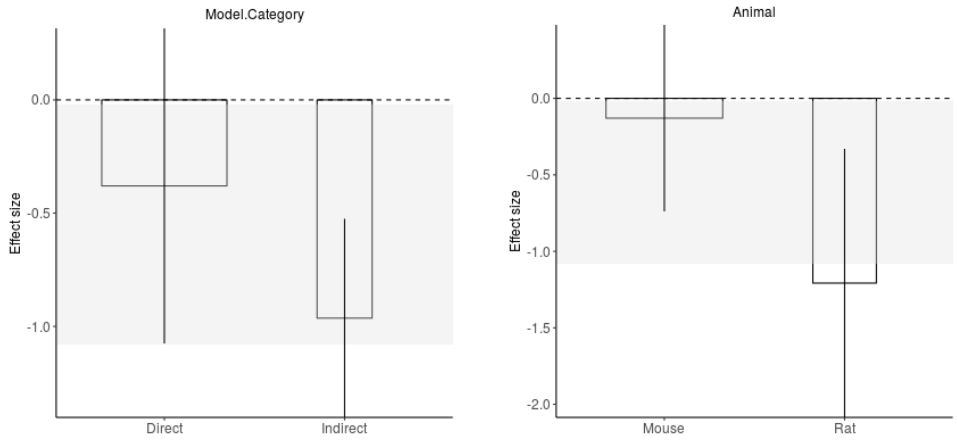


Figure 5.6A & B. (A) Effect sizes associated with model category (4 comparisons reported direct gut-microbiota manipulation, 12 comparisons reporting indirect gut-microbiota manipulation), and (B) type of animal used in the experiment (6 comparisons used rat, 10 comparisons used mice). The grey sections signify the overall effect size including 95% confidence intervals. The bars signify the effect size in each sub-group. Errors bars are 95% confidence intervals.

5.3.3.2 Impact of Measures to Reduce the Risk of Bias

We investigated 4 measures to reduce the risk of bias; random allocation to group, blinded assessment of outcome, allocation concealment, and reporting of a sample size calculation. Random allocation to model or control group did not explain significant heterogeneity between studies. 4 studies reported random allocation to group, 9 studies did not report random allocation to group. No studies reported allocation concealment, and only one study reported a sample size calculation, therefore we were unable to explore these variables as potential sources of

heterogeneity. Only one study did not report a conflict of Interest statement, and compliance with animal welfare regulations. Source of funding did not significantly explain heterogeneity between the studies. 2 studies reported private funding, 5 studies reported public funding, 3 studies reported both public and private funding, 1 study did not report any funding, and 1 study reported no funding. The reporting of measures to reduce the risk of bias are summarised in Appendix 3, Table 4.

5.3.3.3 Other Outcomes

We extracted eight other outcomes from the primary articles; microbial (10 comparisons), neurochemical (58 comparisons), hormonal (9 comparisons), anatomical (8 comparisons), metabolic (3 comparisons), immune (18 comparisons), inflammatory (8 comparisons), and oxidative stress (5 comparisons) outcomes. Due to the small number of studies, these were not able to be investigated with quantitative analysis. An overview of the outcomes investigated in primary articles reporting an intervention to induce depression presented is in Figure 5.7.

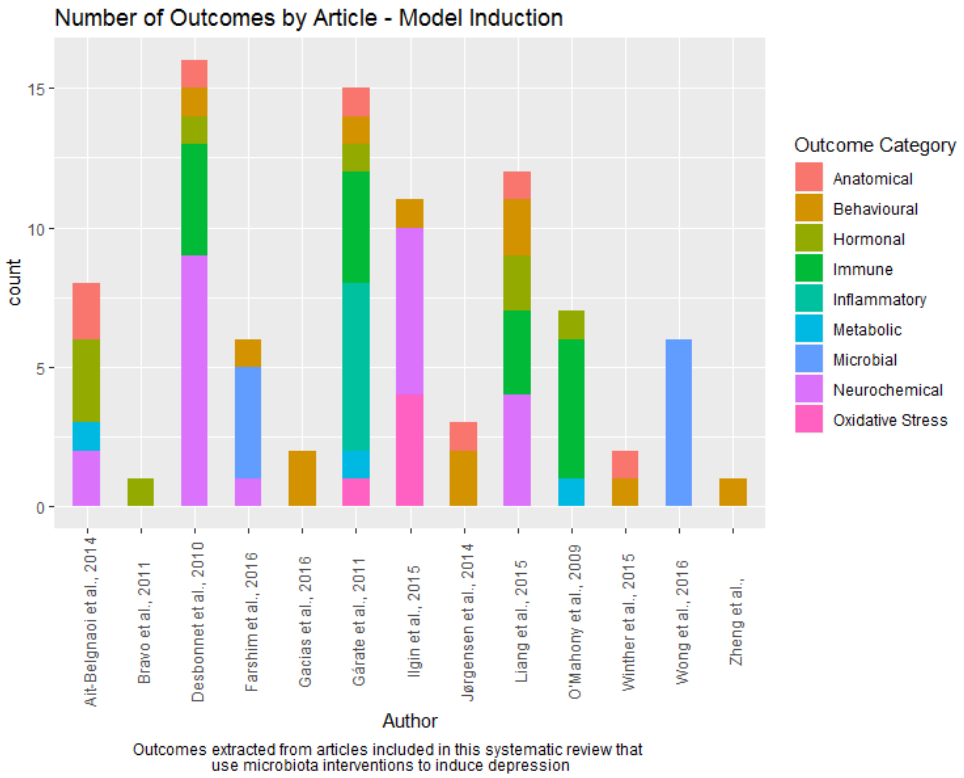


Figure 5.7. The category of outcome measures in experiments investigating gut-microbiota interventions to induce depression. Behavioural (16 comparisons), microbial (10 comparisons), neurochemical (58 comparisons), hormonal (9 comparisons), anatomical (8 comparisons), metabolic (3 comparisons), immune (18 comparisons), inflammatory (8 comparisons), and oxidative stress (5 comparisons) outcomes.

5.3.3.4 Publication Bias

For studies that used gut-microbiota interventions to induce depression, there were sixteen comparisons contributed to the pooled effect size (-0.549 SD; 95% CIs [-1.079; -0.019]). We observed slight funnel plot asymmetry which was not significant with Egger's Regression.

5.3.4 Overview of Neurochemical Outcomes & Brain Regions

There was a broad range of neurochemical outcomes investigated in the primary studies, 67 comparisons in model induction studies and 58 comparisons in treatment

studies. These neurochemical outcomes reported were investigated in many different brain regions. Knowledge regarding, for example, neurotransmitter levels and neurogenesis in depression reports that different neurochemical outcomes are likely to display different levels of expression in separate brain regions. As there were too few studies to investigate the effects of microbiota interventions on individual neurochemical readouts in separate brain regions, this information from all neurochemical outcomes (both model induction and treatment studies) is presented in a diagram to provide an overview (figure 5.8).

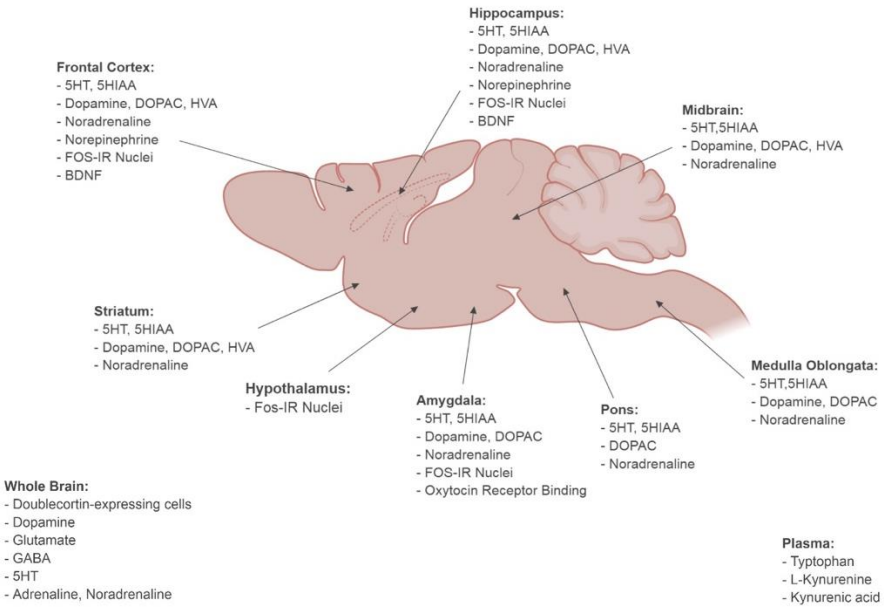


Figure 5.8. Neurochemical outcomes investigated in all studies, both model induction and treatment intervention, separated by brain area they were investigated in.

5.4 Discussion

The results from this systematic review and meta-analysis indicate that up until May 2016, there are few investigations into gut microbiota interventions in animal models of depression. There was a broad range of outcomes investigated in the studies included in this review, from behaviour to inflammatory and oxidative stress response outcomes. Many studies investigated several biological outcomes including behaviour, which can potentially elucidate the underlying biology of depression. The 3Rs advocate for maximising the information gathered from each animal, to reduce the use of additional animals (Prescott & Lidster, 2017).

A range of interventions was used in primary articles. Both antibiotic and probiotic interventions were investigated in models of depression, perhaps reflecting the exploration of mechanisms behind the impact gut microbiota alterations have on mood. No two studies investigated the same probiotic strain; therefore, we cannot compare the relative effectiveness of different strain using meta-analysis. This may reflect the early investigation into these products on depressive-like behaviour in animal models of depression.

Both interventions directly and indirectly targeting the gut microbiota to induce depression were investigated. Indirect models had higher effect sizes than direct models, suggesting that manipulations of the gut microbiota may not induce depressive-like behaviour in animals. Indirect models included traditional model induction techniques such as chronic mild stress. This may reflect the investigation of the impact of stress on the gut microbiota, as a possible additional mechanism through which depressive-like behaviour is induced.

With few studies and many variables investigated, the findings from the meta-analysis of pooled effects of microbiota-targeting treatments and microbiota-targeting model induction methods should be interpreted with caution. This is a rapidly evolving field, where most of the studies included in this review being preliminary studies. An update to this systematic review would ideally allow for data to compare the effects of different strains of probiotics.

5.4.1 Internal Validity

There were 4 studies that reported random allocation to group. This was a significant source of heterogeneity between studies in experiments using microbiota interventions to treat depression, where studies that reported random allocation to control or treatment were associated with higher effect size. This finding should be interpreted with caution as there were only 4 studies contributing to this sub-group in the analysis. Systematic reviews and meta-analyses of animal models in other neurological fields report the opposite to the finding here. Studies that do not report

measures to reduce the risk of bias often have overstated effect sizes (Macleod et al., 2015).

A sample size calculation was only reported in one study. A recent systematic review of the reporting of sample size calculations found that most studies submitted to a leading neuroscience journal, Nature Neuroscience, had lower reporting of sample size calculations and studies were not adequately powered; they did not have adequate sample size to detect even large effects (Carter, Tilling & Munafò, 2017). If studies are not adequately powered, there is an increased probability that significant effects are identified as a false positive.

Blinded assessment of outcome was reported in 40% of treatment studies and in 53% in model studies. This is relatively high in comparison to reviews of animal models in other neurological fields (McCann et al., 2014; Vesterinen et al., 2013; Egan et al., 2016). This may reflect that this is a relatively new field, with papers published since 2010, which was the same year that the ARRIVE guidelines to improve the reporting of animal studies was published (Kilrenny et al., 2010). However, no studies reported all three; randomaisation, blinding, and sample size calculation.

5.4.2 External Validity

Only one study across both meta-analyses used female animals. The sex of animals used in these experiments is consistent with the sex bias in laboratory animals used in biomedical research (Zucker & Beery, 2010). Mental health disorders affect both males and female. In particular for depression, the female to male ratio in disability from depression is 1.7:1 (Murray et al., 2013) and the annual prevalence is 5.5% for females and 3.2% for males (Whiteford et al., 2013). The higher rates of prevalence and disability from the disorder, should be further rationale for investigating the underlying biological effects of stress and the effects of antidepressant treatments in female animals.

The timing of administration is an important factor in the external validity of these experiments and has implications in the use of gut microbiota-targeting treatments. Three studies gave the intervention before the model induction, four studies administered the treatment simultaneously to the model induction, and three studies gave the intervention after the model induction. Typically in the clinic, patients are prescribed an antidepressant after they show symptoms and meet diagnostic criteria for depression. Gut microbiota targeting interventions such as probiotics and prebiotics are available without a prescription, therefore it is possible to take probiotics as prophylactics to reduce the response to a stressful event or taking probiotics during a stressful life-period may improve outcome. Further studies are required to understand if probiotics are able to act as prophylactics and the potential mechanisms behind this.

The most commonly reported behavioural outcome measure was the forced swim test and the tail suspension test. Although, this is widely used as an anti-depressant screening test and used as a measure of depressive-like behaviour, it has been critiqued for being a test of stress coping behaviour or behavioural adaption, and not measuring an internal state of the animal (Molendijk & de Kloet, 2015; Commons et al., 2017). The FST is often used to argue that an experimental design variable has "induced depressive-like behaviour", when it is not a model of depression as the dependant variable is the response to the test rather than an intrinsic state in the animal (Porsolt et al., 1978; Nestler & Hyman, 2010). The predictive validity of current animal models and read-outs may be improved with the use of more naturalistic outcome measures such as assessment of home cage and spontaneous behaviour, and preference testing. These may assist in measuring the intrinsic states of animals, their natural behaviour and their positive or negative emotional states (King, 2003; McArthur & Borsini, 2006), and potentially close the translational gap between studies of animal models of depression and clinical findings.

5.4.3 Publication Bias

There was no evidence of publication bias in this data set. Publication bias is the observation that studies with small or no effects remain unpublished and studies with positive findings are published. Publication bias has been shown to be present in other

fields of preclinical modelling and are seen to have an impact on the statement of efficacy (Macleod et al., 2015; Sena et al., 2010). It is likely that no publication bias was identified as there were a small number of studies included in this systematic review.

5.4.4 Limitations

The main limitation of this systematic review is that the most recent information included was identified in May 2016. The investigation of gut microbiota in animal models of neuropsychiatric disorders is a rapidly evolving field and many studies have been published since. An update to this systematic review is planned, these are only preliminary results. The pre-specified protocol outlined that sex and strain were only to be investigated as potential sources of heterogeneity in model induction meta-analysis and not in treatment analysis. This oversight was not analysed post-hoc as there were few studies.

The protocol pre-specified that all the outcome categories, such as neurochemical and microbial outcomes would have been investigated quantitatively. With the amount of studies included in the systematic review and with the wide range of biological outcomes assessed across different brain and anatomical regions, the decision not to summarise them quantitatively was made. We hope that with an update to this systematic review, there will be more data which can be meaningfully pooled and analysed.

5.5 Conclusion

This systematic review and meta-analysis provides an overview of the literature until 2016 on animal models of depression where gut microbiota-targeting interventions were used. This review found that a broad range of outcomes investigated. Meta-analysis of the primary outcome revealed that gut microbiota-targeting interventions significantly improved behavioural outcomes. Meta-analysis revealed that gut microbiota-targeting interventions to induce depression significantly worsened behavioural outcomes. With few studies and many variables investigated, the findings from the meta-analysis should be interpreted with caution. Microbiota-targeting

interventions are reasonably novel in the field of neuropsychiatry and appear promising. However, it is clear from this study that most of the studies are preliminary studies and further research is needed. A gap in the literature is that interventions were mainly in the form of probiotics or antibiotics, no studies in this systematic review investigated prebiotics. A further gap in the literature is the need for experiments that control threats to interval validity, such as implementing randomisation, blinding, and the conduct of an *a prior* sample size calculation. No studies in this systematic review reported all three of these measures. Therefore, to address these gaps in the literature, I conducted a primary animal experiment investigating the effect of prebiotics in an animal model of depression, the Flinders Sensitive Line, employing measures to reduce the risk of bias. This experiment is reported in Chapter 7. In future, an update to this systematic review would ideally allow for data to compare the effects of different strains of probiotics.

6 THE ANTIDEPRESSANT EFFECT OF KETAMINE IN ANIMAL MODELS OF DEPRESSION AS MEASURED BY THE FORCED SWIM TEST: A SYSTEMATIC REVIEW AND META-ANALYSIS

The findings from this review have been collected and collated with help from Grace Wallace (GW), an honours student at CAMARADES, Oskar Jefsen (OJ), a research year medical student at TNU, Aarhus University, Fraser Sneden (FS), a research assistant at CAMARADES, and Kaitlyn Hair (KH), a PhD student at CAMARADES. Their respective contributions are highlighted throughout. This chapter reports a systematic review and many of the methods are done in duplicate for good practice. Where steps of the review have been done in duplicate, the contribution will be highlighted by using “we/our”.

6.1 Introduction

Major Depressive Disorder (MDD) is the leading source of disability globally (Marcus et al., 2012) and treatment resistance among patients is roughly 50% (Thomas et al., 2013). Therefore, better understanding mechanisms behind MDD and the search for potential effective and novel therapeutic targets are high research and healthcare priorities. Current antidepressant therapies are not rapid acting, the number of patients who experience full remission with citalopram over 8-12 weeks is approximately 33% (Insel, 2006). Treatment resistance, defined as failure to respond to two or more anti-depressant trials (Souery et al., 2006), is common. Even patients that do receive standard care with treatment of SSRIs, only respond adequately approximately 30-50% of the time (Rush et al., 2006; Thomas et al., 2013). Conway and colleagues note that current definitions are vague (ranging from 1 to 8 failed antidepressant treatment trials), which has likely contributed to the varying estimate of prevalence of treatment resistant depression (Conway et al., 2017). They propose an update to the operational definition of treatment resistant depression (Conway et al., 2017), to include a fixed number of anti-depressant trials the patient has tried and call for a definition where antidepressants with different mechanisms of action have been tested. Despite work being carried out to correctly identify treatment resistant depression, it is clear that the current treatments are not adequate for all patients with

depression, and where current treatments are effective they are not rapid-acting to reduce symptoms.

Animal models are commonly used to mimic aspects of the phenotype of the human disorder and to characterise candidate antidepressant agents. Animal models have been established to mimic treatment-resistant depression. Treatment resistant depression is defined in animal models as not responding to clinically used antidepressants and are validated by comparing treatment resistance rates in patients (Caldarone et al., 2015). Caldarone and colleagues review the commonly used animal models for mimicking treatment resistant depression, which include administration of hormonal or inflammatory agents such as ACTH and IL-6, stress models such as CMUS with non-responder groups, genetic models such as BALB/cOLaHsd as well as single gene transgenic animals, and combinations of these models (Caldarone et al., 2015). ACTH administration appears to be the most commonly used model. A dose of approximately 100ug ACTH, when administered for 14 days has been shown to be most effective (Kitamura et al., 2002), and this finding has been replicated in rats and mice (Walker et al., 2013 (rats); Caldarone & Brunner, 2009; Iwai et al., 2013 (mice)).

One potential target for treatment resistant depression is ketamine. Ketamine, the N-methyl-D-aspartate (NMDA)-receptor antagonist, initially marketed in 1970, has been used across healthcare and veterinary medicine for anaesthesia. More recently, it was discovered that ketamine could reverse depressive-symptoms in patients with treatment resistant depression (Berman et al., 2000). Since this initial randomised controlled trial, several subsequent trials have been conducted. Seventy-three studies were retrieved in Clinicaltrials.gov when searching ketamine and major depressive disorder (August 2018), twenty-seven of which are completed. In the published reports from these clinical trials, the anti-depressant response is rapidly acting, within 2-4 hours, which is maintained for 4-7 days (Williams & Schatzberg, 2016). Approximately 40-60% of patients showed reduced symptoms after 24 hours (Williams & Schatzberg, 2016).

The mechanisms of action that ketamine have been found to act through include synaptic plasticity and neurotrophic signalling (Murrough, 2012). Ketamine has been shown to increase the number of synapses and the synaptic function in the PFC through mammalian target of rapamycin (mTOR) and extracellular signal-regulated kinase (ERK) signalling pathways (Liu et al., 2017). The antidepressant activity of ketamine has been shown to be dependent on rapid synthesis of BDNF through tyrosine kinase receptor (TrkB) (Autry et al 2011). Ketamine has been shown to decrease activity in the PFC and the orbital frontal cortex and increase activity in the posterior cingulate in a human fMRI study (Deakin et al., 2008). The behavioural and biological effects of ketamine have been investigated in standard animal models of depression as well as animal models of treatment resistant depression.

The forced swim test (FST) is a commonly used assessment tool to test anti-depressant efficacy and assess depressive-like behaviour. The FST was developed by Porsolt and colleagues in 1977. The FST involves subjecting animals to an acute stressor of being in a water tank from which they are unable to escape, where animals after some time give up and display immobility or behavioural despair. The FST is used in mice and rats with slight modifications between the species. When the FST is used with rats, they are often subjected to a pre-swim session, with no pre-swim session used typically in studies involving mice.

Systematic review and meta-analyses of non-human animal data are hypothesis-generating tools which provide an overview of the field and can be used to inform the experimental design. Systematic reviews in other neuroscientific fields, such as Alzheimer's disease, stroke, and multiple sclerosis, have provided evidence as to the components of experimental design that may influence efficacy of an intervention (Egan et al., 2016; O'Collins et al., 2006; Vesterinen et al., 2010). We aim to apply these methodologies to the use of ketamine in animal models of depression and understand the effects on depressive-like behaviour.

6.1.1 Aim

In this systematic review and meta-analysis we set out to review the efficacy of ketamine in animal models of depression that assess depressive-like behaviour on the forced swim test (FST). We aimed to provide an overview of the literature. We aimed to describe the quality of the data available, and if possible, meta-analysis to investigate study design characteristics that influence efficacy (dose, timing, and route of administration of ketamine).

6.2 Methods

The methodology for this systematic review was laid out in a pre-specified protocol published on the CAMARADES repository of Systematic Review Protocols (http://www.dcn.ed.ac.uk/camarades/research.html#protocols) on 8th February 2018.

6.2.1 Search Strategy

Studies reporting the use of ketamine in animal models of depression were identified from an existing database of studies of animal models of depression. This database was collated using the search string in Bannach-Brown et al., 2016 (Chapter 2). The original search was carried out in May 2016. This database was searched using the regular expression dictionary terms; [kK]etamine|[kK]etalar|[eE]sketamine|JNJ-54135419|[kK]etanest S|\([sS]\)[-|]?[kK]etamine|[sS]\(\+\)[-|]?[kK]etamine

Abstract screening was carried out by GW and ABB, any screening discrepancies were resolved by KH. Full-text screening was carried out by GW & OJ to identify papers that reported forced swim test outcomes. GW, OJ, FS and ABB carried out full-text screening and data extraction. KH performed reconciliation of outcome data.

6.2.2 Inclusion & Exclusion Criteria

Publications were included if they tested ketamine in the Forced Swim Test (FST) to measure depressive-like behaviour, in an *in vivo* experiment, where any quantitative outcome of depression had been assessed. Studies were included if there was an

appropriate control group (received a vehicle of saline in the same volume) and where the sample size, the mean and the variance in each group had been reported for the primary outcome, behaviour. There were no exclusion criteria based on species or strain of animal, age, sex, weight or method of model induction. There were no exclusion criteria based on dosage, timing or frequency of administration of the ketamine. Studies were excluded if only anxiety-like behaviour was the only behavioural outcomes investigated. There were no exclusion criteria pertaining to date of publication but studies were excluded at full-text if not in English due to the time constraints of the project.

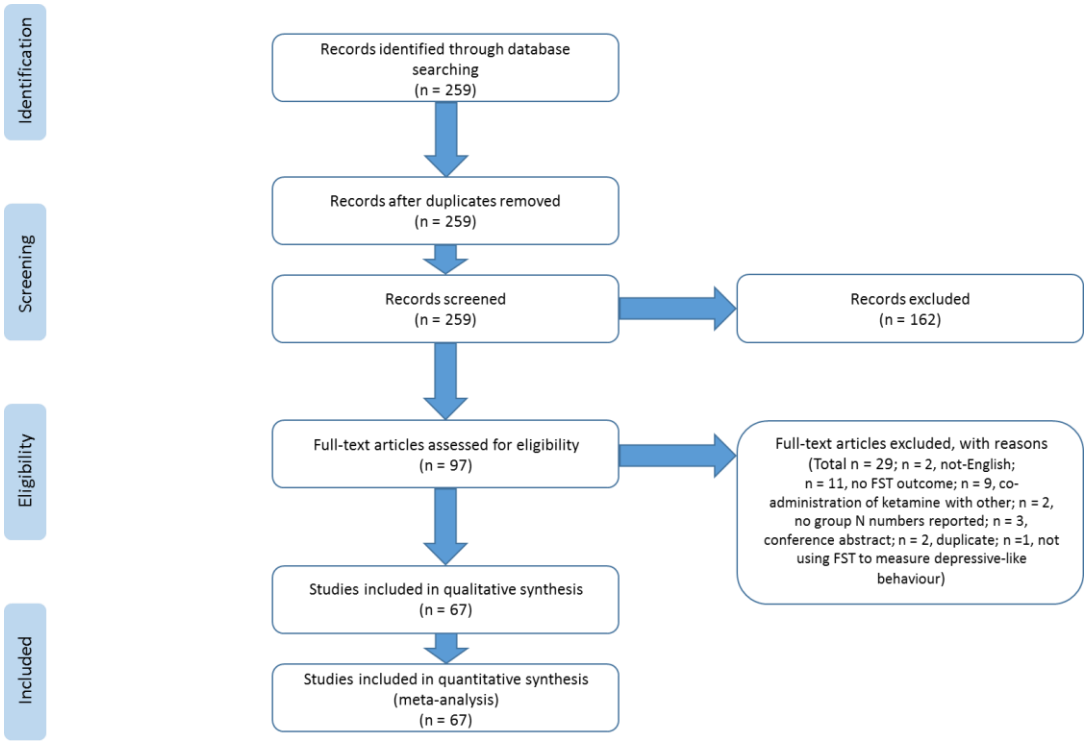


Figure 6.1. PRISMA flowchart of studies included in this review.

6.2.3 Data Extraction from Primary Studies

For each publication, information on quality (see the following section) and experimental design characteristics (animal species, strain and intervention tested) were entered into the SyRF, an online Systematic Review Facility developed by CAMARADES & NC3Rs. For all included studies, details regarding the following information was extracted:

- The method of model induction, such as stress, pharmacological or lesion models
- The species, strain, sex and age or weight of animals
- Ketamine characteristics:
    - Type of ketamine (Ketamine, R-ketamine, S-ketamine)
    - Dose
    - Route of administration
    - Number of administrations
- Age of animal at administration (if reported)
- Time between ketamine administration and outcome assessment
- Animal husbandry conditions

The number of animals, mean and measure of variance (either SD or SEM) for control group and treatment group were extracted for each outcome of interest. The primary outcome measure was efficacy in the forced swim test. Any measure from the forced swim test was extracted (e.g. immobility, swimming, struggling, or head-shakes)

Where data were presented graphically, universal desktop ruler (https://avpsoft.com/products/udruler/) or the in-built measuring tool in Adobe was used to extract numerical values. Data extraction was carried out in duplicate by GW, FS, OJ and ABB. KH carried out reconciliation of extracted outcome data.

6.2.4 Assessment of Quality

Methodological quality was assessed by recording the reporting on any of the following four items; whether a study reported the randomisation of animals to control and treatment groups, whether a study reports the blinding of investigators under the assessment of outcome measures, allocation concealment, and if authors report a sample size calculation (Macleod et al., 2004). Further, we assessed whether authors report that experiments comply with animal welfare regulations, whether authors have included a statement of potential conflicts of interest, and what the source of funding was, if reported.

6.2.5 Data Reconciliation

Two independent reviewers (GW, OJ, FS & ABB) extracted data from publications. Extracted data at the level of the publication and the outcome were compared and any discrepancies were checked and corrected. Effect sizes for each comparison were calculated for both reviewers and compared, where effect sizes differed by more than 10%, the original article was checked and data was re-extracted from graphs or tables in the primary articles. Where effect sizes for comparisons differed by less than 10%, the mean of the effect sizes and the errors (SEM/SD) were calculated.

6.2.6 Data & Meta-Analysis

For the primary outcome, the forced swim test, the effect size measure standardised mean difference (SMD) was calculated for each comparison. As above, a comparison is defined as a pair of cohorts of animals, one cohort that receives an intervention, and the other cohort receives an appropriate control for that intervention (Vesterinent et al., 2014). SMD was used as we were not able to infer the 'normal' behaviour or 'normal' biological readouts of an animal, and data are presented on different scales. All equations used are taken from Vesterinen and colleagues (2014) and are described fully in Chapter 5. The effect sizes from each comparison were combined using random-effects modelling, with Hartung-Knapp adjustment, with a restricted maximum likelihood estimate of between study variance. Random effects meta-analysis assumes that true effects differ between studies and allows us to explore the differences between the studies that contribute to differences in effect sizes. If several measures were taken from the FST, e.g. immobility, swimming, and struggling, were reported from the same cohort of animals, these were combined or nested using fixed effects meta-analysis and this aggregate estimate was used for further analysis in the random-effects model. Fixed effects meta-analysis is similar to random effect meta-analysis. However, the assumption is that true effects do not differ between comparisons. The weight of each study is the inverse of the variance. Where multiple intervention groups were used with a single control/vehicle group, the number of animals in the control group was divided by the number of intervention groups. Random-effects meta-analysis and meta-regression was carried out in Stata using the metareg function. Nesting was carried out in R. The impact of study design characteristics and study quality was assessed with meta-regression. The study design and intervention characteristics pre-specified are outlined in Table 6.1 below.

Table 6.1 Pre-specified study design characteristics.

Study Design Characteristics investigated:
Method of Model Induction (Chronic Stress, FST, genetic, learned helplessness, CORT/LPS insult, Other)
Sex of animals
Age or weight of animals
Species of animal
Form of ketamine (Ketamine, R-ketamine, S-ketamine, DehydroNorKetamine, Ketalar, NorKetamine)
Dose
Route of administration
Frequency of administrations (Single admin or multiple admins)
Timing of treatment administration (before, during, after model induction)
Timing of treatment in relation to outcome assessment

In table 6.1, the characteristics not underlined were unable to be investigated. The age of the animals was not able to be investigated as a variable of interest, because primary articles reported either age in weeks or weight in grams, which were unable to be combined validly. No studies reported a sample size calculation, so this was unable to be investigated. Only one study investigated an alternative route of administration, other than intraperitoneal, therefore this was not investigated. Meta-regression was used to investigate treatment dose as a possible source of heterogeneity. Meta-regression was performed using a Meta-analysis Online Platform based on R (code available here: https://github.com/qianyingw/meta-analysis-app). Methodology behind this approach is outlined in (Chapter 5: Methods > Data & Meta-Analysis).

A simulation study by Wang et al., (2018) at CAMARADES identified that stratified meta-analysis using SMD has low statistical power to detect the effect of a variable of interest, but a low false positive rate, so we can have confidence in any significant results found. A Holm-Bonferroni correction was applied to correct for multiple testing. The significance level for the analysis of study design in ketamine experiments was adjusted to $p < 0.006$. The significance level for the analysis of measures to reduce the risk of bias in ketamine experiments was adjusted to $p < 0.009$.

6.2.7 Publication bias

The risk of publication bias was assessed using visual inspection of a funnel plot, Egger's regression (Egger et al., 2007) and trim and fill analyses (Duval & Tweedie, 2000) were used to identify potentially missing studies.

6.3 Results

6.3.1 Identifying publications

A systematic search conducted in PubMed and EMBASE in May, 2016 retrieved 70,365 unique publications. Machine-learning algorithms were employed to screen the studies based on a sub-set of documents with human decisions (2 independent reviewers with 3rd for reconciliation). A performance of 98.7% sensitivity and 88.7% specificity was achieved (for more detail of the methodology used, see Chapter 3). 18,409 documents were included in the review by the machine-assisted approach, which formed the depression database. 259 studies were identified in the database using the Regex Dictionaries. These 259 studies were screened by two independent human reviewers with a third reviewer resolving any disagreements. 97 articles were included at the title and abstract screening stage. 67 publications that reported FST outcomes were included in this review after full-text screening. The most frequent reason for exclusion at the full-text screening stage were publications not in English, or publications not reporting the FST as an outcome. The study selection process is highlighted in the PRISMA flowchart in figure 6.1.

6.3.2 Effect of ketamine on depressive-like behaviour in the FST

The effect of ketamine on behaviours in the forced swim test was reported in 67 publications. After nesting the instances where several outcomes from the FST were reported in the same cohort of animals at the same time of assessment, we had 182 comparisons, using 3,203 animals. Data are summarised in Appendix 3 Tables 1 and 2.

The use of ketamine was reported in 159 experiments, 6 reported Ketalar, 5 reported S-Ketamine, 5 reported NorKetamine, 4 reported R-Ketamine, and 3 comparisons

reported DehydroNorKetamine. 85 experiments reported administering ketamine before the model induction, 18 reported administration during model induction, and 79 comparisons reported administration after model induction. 98 experiments reported using mice, 84 reported using rats. 151 comparisons reported using male animals, 18 using female animals, 7 using both sexes, and sex was not reporting in 7 comparisons. Sample size of treatment groups ranged from 5 animals to 18 animals, with a median of 8 animals per group. The dose administered ranged from 0.25 mg/kg to 160 mg/kg. The timing between the treatment administration and outcome assessment ranged from 19 days before outcome assessment to 20mins prior to outcome assessment. Ketamine was reportedly administered once in 150 experiments, and ketamine was administered more than once in 32 studies. The method of model induction was an immune or inflammatory insult (corticosterone, dexamethasone, lipopolysaccharide, or adrenocorticotropic hormone) in 13 experiments, chronic stress was administered in 49 experiments, a genetic model of depression was used in 12 experiments, the forced swim was the only induction of stress in 101 experiments, learned helplessness was used in 5 experiments, and 2 utilised other methods of model induction (drug-withdrawal and neuropathic pain induced depression). This data is summarised in Appendix Table 1. Only one study investigated an alternative route of administration, other than intraperitoneal, therefore this was not investigated using meta-analysis.

Taking all forms and doses of ketamine together, there was a significant improvement in depressive-like behaviour in the FST (0.869 SD, 95% CIs [0.738; 1.001], figure 6.2. There was no significant statistical evidence of heterogeneity observed between the comparisons (τ^2= 0, I^2 = 0.0%, Q = 78.8, df = 181, p = 1).

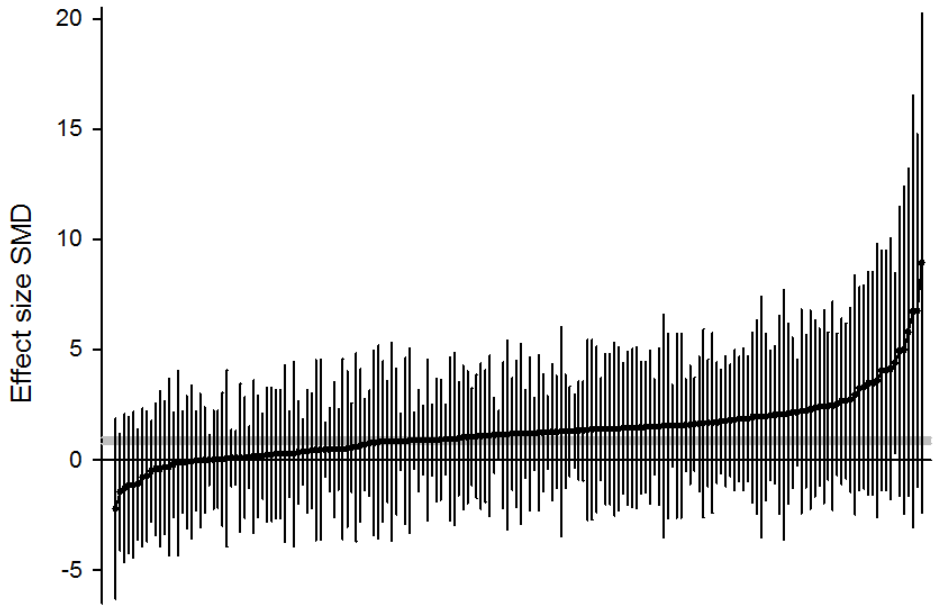


Figure 6.2. Timber plot of the effect sizes from 182 experiments investigating the effect of ketamine in forced swim test outcomes. Black dots represent effect size (SMD), error bars represent 95% confidence intervals. The grey bar behind the plot represents the 95% confidence intervals of the pooled effect size [0.738; 1.001].

6.3.3 Publication Bias

I observed funnel plot asymmetry, which was confirmed with Egger's regression. Trim and fill analysis suggested 43 theoretical missing studies correcting the SMD effect size to 0.685 SD [0.497; 0.873] when these missing studies were included in pooled effect size (see figure 6.3). This suggests a 27% relative overestimation in treatment effect, an absolute difference of 0.184 SD.

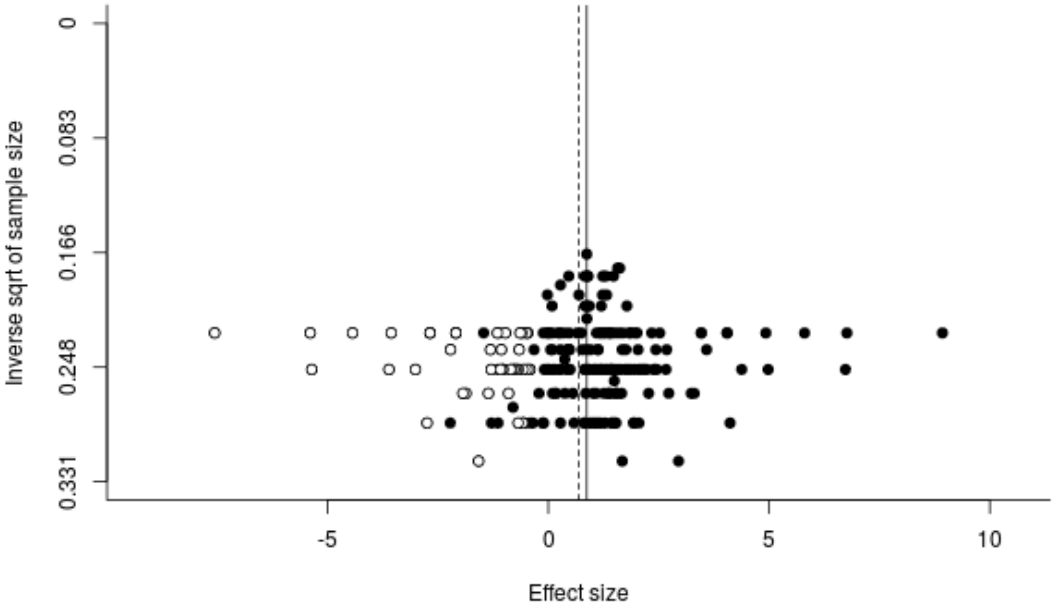


Figure 6.3. In experiments reporting the effects ketamine on depressive-like behaviour, plotting studies by effect size and the inverse square root of the sample size on a funnel plot suggests asymmetry. Filled black circles represent published experiment. Unfilled white circles represent theoretically missing studies imputed with trim and fill analysis. The solid line represents the global effect size (0.869 SD) and the dashed line represents the adjusted global effect size which includes the theoretically missing studies (0.685 SD).

6.3.4 Impact of Study Design Variables

Planned analysis of the effects of study design variables on heterogeneity was not carried out as no heterogeneity was observed in this dataset, however, I did assess

whether these variables were associated with effect size. The study design and intervention characteristics (method of model induction, sex of animals, species, the form of ketamine, dose, frequency of administration, timing of treatment administration, and timing of treatment in relation to outcome assessment) were investigated using a multivariate meta-regression model with Knapp-Hartung modification. The timing of administration was significantly associated with a reduction in effect size (beta coefficient: -0.004 [-0.006; -0.002], p = 0.001, figure 6.4), when other variables were kept constant. For every 1 minute prior to the outcome assessment, there was a -0.004 decrease in depressive-like behaviour. None of the other study characteristic variables had a significant association with the effect size. A summary of these study design characteristics is available in Appendix Table 1.

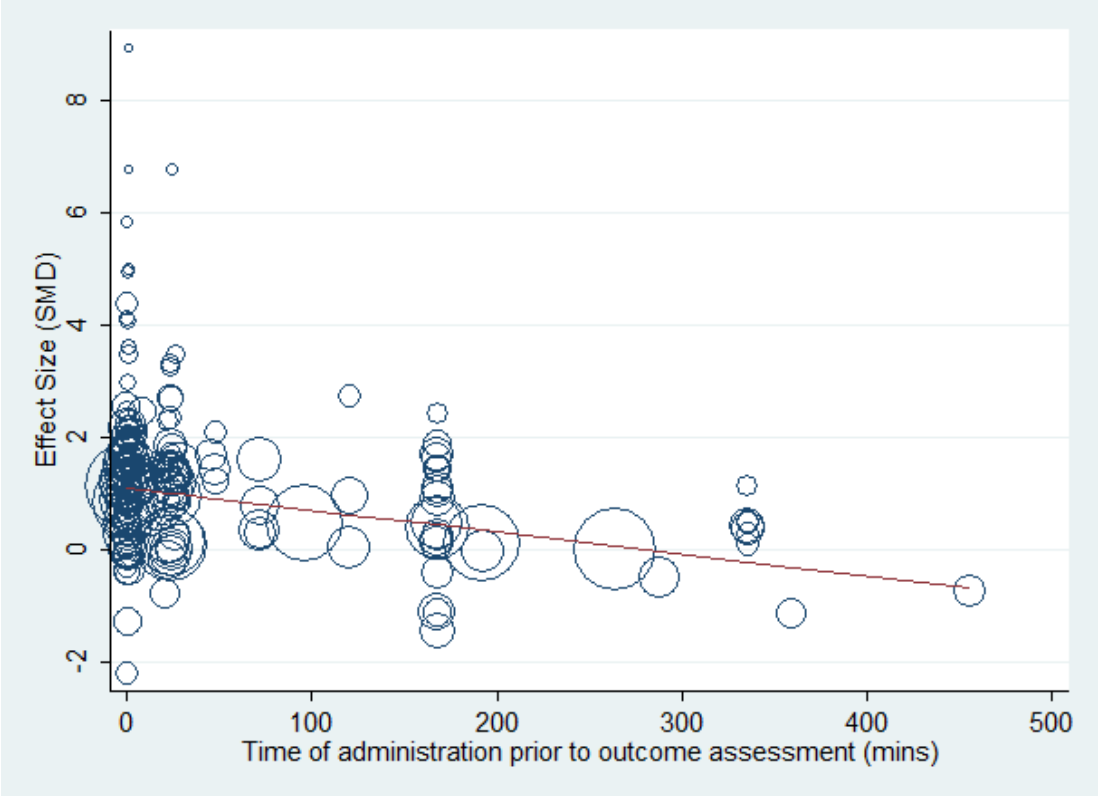


Figure 6.4. A visualisation of the effect of timing of administration of ketamine (mins prior to outcome assessment) on effect size (SMD). Circles represent the effect size of individual comparisons. The size of the circle represents the precision of the study (the inverse of its within-study variance). The red solid line represents a linear representation of the prediction of effect size as a function of the dose.

6.3.5 Impact of Measures to Reduce the Risk of Bias

Planned analysis of the effects of measures to reduce the risk of bias on heterogeneity was not carried out as no heterogeneity was observed in this dataset. However, I assessed whether these measures were associated with effect size. The impact of reporting study quality characteristics (random allocation to group, blinded assessment of outcome, allocation concealment, conflicts of interest, compliance with animal welfare regulations, and source of funding) were investigated using a multivariate meta-regression model with Knapp-Hartung modification. None of the study characteristic variables had a significant association with the effect size at the corrected alpha level. 18% of studies reported randomisation of animals to group. 38% of studies reported blinding assessment of outcome. 3% of studies reported allocation concealment and 0% of studies reported conducting a sample size calculation. An overview of the reporting of measures to reduce the risk of bias for each study is presented in Appendix Table 2.

6.3.6 Other Behavioural Outcomes

We extracted the reporting of other behavioural outcomes from the primary articles. Other than the forced swim test, 37 studies reported depressive-like outcome assessments (tail suspension test, sucrose preferences test, and learned helplessness), 32 reported the open field test, 21 studies reported assessment of anxiety-like behaviour (novelty suppressed feeding or hypophagia, elevated plus maze, marble burying, light-dark box, and the hole board test), 10 studies reported assessment of locomotor activity on the rotarod or other equipment, 5 studies reported the assessment of cognitive outcomes (attentional set-shifting or conditioned place preference), and 8 studies reported the assessment of other behaviour outcomes (the splash test, cold or mechanical allodynia, prepulse inhibition, and stereotype rating). See figure 6.5.

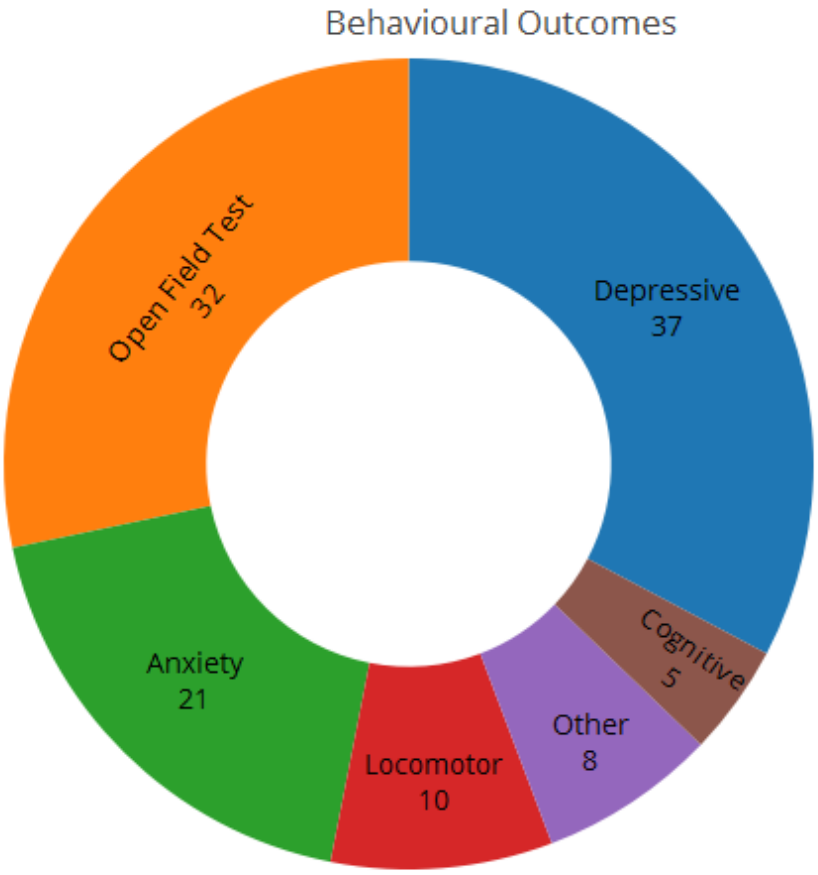


Figure 6.5. A doughnut plot of the additional behavioural outcomes assessed in studies investigating ketamine in the forced swim test. Sections of the doughnut plot represent the number of studies reporting the category of behavioural outcome.

6.4 Discussion

This study assesses the preclinical literature reporting administration of ketamine for the treatment of depressive-like behaviour in animals. We investigated the effects of ketamine on behaviour in the forced swim test. Pooling data from 67 studies, ketamine significantly improved depressive-like behaviour in the forced swim test. We did not observe any statistically significant heterogeneity between the studies. The time ketamine was administered before the outcome assessment had a significant negative association with effect size, with longer time between administration and outcome assessment associated with lower effect sizes.

Interestingly, the 95% confidence intervals for individual effect sizes crossed zero. With pooling the data together into a summary effect, this increases the precision of the estimate. The lack of statistically significant heterogeneity between studies may be due to the imprecision of the individual studies. In this meta-analysis, effect sizes ranged from -2.22 SD to 8.97 SD. However, with overlapping confidence intervals, the relative differences in effect size are not statistically identified as heterogeneous. Further, if the true heterogeneity between the studies is low, many studies are required to show significant heterogeneity between the studies. Further, tests to detect heterogeneity are generally low powered (Higgins et al., 2003). The forced swim test is a uniform test with many researcher s following the original protocol by Porsolt and colleagues (1978). There may be slight differences between experiments, for example, the temperature of the water, the length of the test session, the use of a pre-swim session, testing in the light or dark phase of the animals, which have not been investigated in this review.

Several forms of ketamine were investigated in this review. Ketamine was investigated in 87.3% of experiments, 6% of experiments reported investigating the enantiomers S-ketamine and R-ketamine. Although the form of ketamine used was not significantly associated with effect size, the investigation of different enantiomers in animal models may provide useful understanding of the potential side-effects and inform clinical trial design (Muller et al., 2016). The smallest dose administered was 0.25 mg.kg and the highest doses administered by primary studies in this review were 160 mg/kg, which may have neurodegenerative effects (Green & Cote, 2009).

6.4.1 Internal validity

18% of studies reported randomisation of animals to group. 38% of studies reported blinding assessment of outcome. 3% of studies reported allocation concealment. This is generally poor reporting of measures to reduce the risk of bias. A sample size calculation was not reported in any study. When studies are not adequately powered there is an increased probability that significant effects are false positive (Carter, Tilling & Munafò, 2017). No single measure to reduce the risk of bias was significantly associated with effect size when all other risk of bias variable were accounted for. We

were not able to investigate the impact of measures to reduce the risk of bias on heterogeneity between studies.

6.4.2 External validity

The timing of administration is an important factor for the external validity of the findings. 46% of experiments reported administering ketamine prior to model induction (85/182 comparisons), which may not reflect the treatment regime in clinical settings. Typically in the clinic, patients are prescribed an antidepressant treatment after they show symptoms and meet diagnostic criteria for depression. Ketamine is being investigated as a rapid-acting antidepressant in treatment-resistant depression and in acute instances of suicidal ideation (Kashani et al., 2014). Therefore, the external validity of findings from studies that administer ketamine prior to model induction may be brought into question. The external validity is called into question even if, as in this case, no significant differences in effect sizes were observed between studies that report administration before or after model induction. This non-significant finding may have been affected by the observed imprecision of primary studies.

As with my systematic review of microbiota-targeting interventions (Chapter 5), sex bias is obvesrved in the animals used in primary studies of ketamine efficacy. 18 out of the 182 experiments reported using female animals, and 7 out of 182 experiments reported using both sexes, which is unfortunately consistent with experiments in other areas of biomedical research (Zucker & Beery, 2010). This highlights a gap in the research. Further research can be conducted to understand the mechanisms responsible differences between sexes in animal models of depression to inform human treatment.

6.4.3 Publication Bias

There was evidence of publication bias in the literature reporting the effects of ketamine in animal models of depression. Trim and fill analyses imputed 43 hypothetical studies. Egger's regression estimated that there was a 27% relative overstatement of efficacy of ketamine in this literature. This suggests that studies with

smaller or negative effects are less likely to be published in this field. This result is similar to findings of publication bias has been identified in other neurological fields (Sena et al., 2010).

6.4.4 Limitations

The main limitation of this systematic review is that the search was carried out in May 2016. Several studies may have been published since which may alter the pooled effect size estimate. Further, this systematic review only reports data from one outcome, the forced swim test. The analysis of other depressive-outcomes, such as the tail suspension test and sucrose preference test, as well as other behavioural outcomes, such as cognitive and anxiety-like behaviour, may add to our understanding of the variables that contribute to any differences in effect sizes between studies, if these tests are more precise than the forced swim test. The investigation of other behavioural outcomes may help shed light on potential side-effects such as hyperlocomotion and systematic review of this data may allow for analysing the effects of dose on side-effects.

6.5 Conclusion

This systematic review and meta-analysis provides an overview of the literature on animal models of depression that investigate the effects of ketamine in the forced swim test. Overall, ketamine significantly reduced depressive-like behaviour. The timing between ketamine administration and outcome assessment was associated with significant decreases in effect size. There was no statistically significant heterogeneity in this dataset. This may be due to the low power of heterogeneity tests or the imprecision of the primary studies included in the review. There was low reporting of measures to reduce the risk of bias in the literature reviewed. Study design characteristics, such as 46% of studies administering ketamine prior to model induction, only 18 studies using female animals, and all but one study administering ketamine intraperitoneally, call into question the external validity of the studies. Further animal studies are needed that explore aspects of experimental design that can improve the translation of findings from animal studies into clinical trials. Further investigation into behavioural outcomes such as anxiety-like behaviour and locomotor activity may provide insight into the potential side-effects of ketamine that have been

investigated with animal models of depression and allow for the investigation of heterogeneity if these experiments are more precise.

7 ADMINISTRATION OF GALACTO-OLIGOSACCHARIDE PREBIOTICS IN THE FLINDERS SENSITIVE LINE ANIMAL MODEL OF DEPRESSION

The experiment in this chapter was designed with help from Sandra Tillmann, a PhD student at the Translational Neuropsychiatry Unit, Aarhus University and Professor Gregers Wegener.

7.1 Introduction

Major Depressive Disorder (MDD) is the leading source of disability globally (Marcus et al., 2012) and treatment resistance among patients is roughly 50% (Thomas et al., 2013). Therefore, better understanding mechanisms behind MDD and the search for potential effective and novel therapeutic targets are high research and healthcare priorities. Animal models are commonly used to mimic aspects of the phenotype of the human disorder and to characterise candidate antidepressant agents. The Flinders Sensitive Line (FSL) is a well-established and validated genetic model of depression (Overstreet & Wegener, 2013). The FSL rats are bred to display cholinergic sensitivity and later found to display depressive-like behaviour in the forced swim test (FST), compared to their control strain, the Flinders Resistant Line rats (FRL) (Overstreet & Wegener, 2013). FSL rats respond to acute and chronic antidepressant administration and display reduced hippocampal plasticity (Chen et al., 2010) and elevated rapid eye movement (REM) sleep (Benca et al., 1996) in comparison to FRL rats.

Communication between the gut microbiome and the brain may play a role in psychiatric disorders, with research focussing on the bidirectional signalling at the neural, hormonal and immunological levels (Cryan & O'Mahony, 2011). Interventions targeting the gut microbiota may serve as potential treatments for depression, and this drives increasing research into the effect of probiotics and prebiotics in neuropsychiatric disorders. Probiotics have been defined as "live organisms, that when ingested in adequate amounts, exert health benefits." (Dinan et al., 2013). Several probiotic strains have been investigated in psychiatric disorders and have reported effects on behaviour and physiology, in laboratory animals and humans (for a review see Wang and colleagues (2016)). Commercially available probiotic products, "Ecological Barrier" and "Probio'Stick", have been tested in FSL rats (Abildgaard et

al., 2017; Tillmann, et al., 2018). Prebiotics, defined as galacto-oligosaccharides and fructo-oligosaccharides that stimulate the activity of *Bifidobacteria* and *Lactobacilli* in the gut, have also been reported to have a positive impact, reducing anxiety and depressive-like phenotypes and stress-related physiology in mice and rats (Burokas et al., 2017; Mika et al., 2017; McVey Neufeld et al., 2017; Thompson et al., 2017; Savignac et al., 2013) and in humans (Perez-Cornago et al., 2016; Schmidt et al., 2015; Tillisch et al., 2013). Further, prebiotics have been shown to increase the diversity of gut microbial composition, with evidence from mice (Burokas et al., 2017) and rats (Mika et al., 2017). Thompson and colleagues (2017), however, showed no difference in gut microbiota composition in F344 rats receiving prebiotics. One prebiotic that is commercially available is Bimuno®. Bimuno® contains beta-galactooligasaccharide (B-GOS) produced from lactose in cow's milk (Bimuno, 2018).

In Chapter 5 I conducted a systematic review and meta-analysis of literature reporting gut microbiota-targeting interventions in animal models of depression and identified gaps in the current literature. The systematic reivew identified that interventions used were mainly in the form of probiotics or antibiotics, no studies in this systematic review investigated prebiotics. A further gap in the literature was the need for experiments that control threats to interval validity, such as implementing randomisation, blinding, and the conduct of an *a prior* sample size calculation. No studies in the systematic review reported all three of these measures.

Based on previous studies showing promising effects of other prebiotics to reduce depressive- and anxiety-like behaviour and based on a systematic review of literature reporting gut microbiota-targeting interventions (Chapter 5), we will investigate the effect of Bimuno® on rat behaviour and gut microbiota composition in the FSL model, a genetic model of depression, in comparison to their control FRL rats.

7.1.1 Hypotheses

We hypothesise that FSL animals receiving Bimuno® prebiotics will display reduced depressive-like behaviour in the FST and reduced anxiety-like behaviour in the Elevated Plus Maze (EPM) in comparison to control (same composition but without active ingredient). As our secondary outcome, we hypothesise that FSL animals receiving prebiotics will display increased diversity in the gut microbiome, in comparison to FSL animals receiving control, as measured on true beta diversity. We want to analyse gut microbiome diversity because we hypothesise that this is the mechanism through which prebiotics influence behaviour. Little is known about these mechanisms and we therefore aim to shed light on the commensal influence of prebiotics (Sherwin et al., 2016). We hypothesise that animals receiving Bimuno® prebiotics will have altered weight and food intake in comparison to animals receiving control.

7.2 Methods

7.2.1 Animals

Ethics has been approved by Aarhus University animal ethics committee (permission ID 2012-15-2934-00254). 8-14 week-old male FSL and FRL rats bred in-house at TNU, Aarhus University were used. Animals were housed in pairs, in standard cages with a plastic bottom and metal rack top half, purchased from Techniplast (Cage 1291H Eurostandard Type III H, 425 × 266 × 185 mm, Tecniplast, Italy). The bedding material in each cage was made out of wooden chips (aspen wood from Tapvei®, Finland) along with access to a tunnel shelter, nesting material, and a wooden stick. Animals were maintained in a 12hr light/dark cycle with lights off at 1300hrs (time point 0). Seven days prior to the experiment start, the animals were moved to the experimental facility and the new lightning regime was started. Animals were under the care of FELASA-accredited in-house animal technicians. Animals had free access to tap water and standard chow (purchased from Brogaarden®, Altomen 1324).

7.2.2 Power calculation to determine the number of animals

Our sample size calculations are based on published behavioural findings from Burokas and colleagues (2017) and McVey Neufeld and colleagues (2017).

Data were extracted from Burokas et al., (2017) who investigated effects of prebiotics in the FST using male C57L/6J mice (Fig. 6D in the publication). These data (mean, SEM, and group numbers) were used to run a one-way ANOVA and determine an eta squared (= SSbetween/SStotal) of 0.579. This eta squared value was used to compute effect size f = ($\sqrt{\left({eta^2}/{1 - eta^2}\right)}$) which is 1.1747. This effect size was used in the power calculation carried out in R (v. 3.4.3) using the function "pwr.anova.test". A significance level of 0.01 and a power of 0.9 were chosen. This gave the result of 6 experimental units per group. For full R code see Appendix 5. An experimental unit is the entity subjected to an intervention independently of all other units where it is possible to assign two experimental units to different treatments groups (NC3Rs, 2018). In this experiment, the cage is the experimental unit.

Data were extracted from McVey Neufeld et al., (2017) who used prebiotics and probiotics in a maternal separation model of depression in the open field using male Sprague-Dawley rats. Data is from the amount of time spent in the centre of the open field (Fig 1.B in the publication) for the model group. These data (mean, SD or SEM, and group numbers) were used to run a one-way ANOVA and determine an eta squared (= SSbetween/SStotal) of 0.522. This eta squared value was used to compute effect size f = ($\sqrt{\left({eta^2}/{1 - eta^2}\right)}$) which is 1.046. This effect size was used in the power calculation carried out in R (v. 3.4.3) using the function "pwr.anova.test". A significance level of 0.01 and a power of 0.9 were chosen. This gave the result of 6 experimental units per group. For full R code see Appendix 5.

Based on the a priori sample size calculations above and experience from previous in-house experiments, a conservative estimate of sample size for this study of 8 experimental units per group was selected. This number is two per group larger than the power calculation and was selected to account for possible attrition or possible exclusions throughout the experiment (see criteria below). With 8 experimental units per group, power of 90%, and a significance level of 0.01, we are powered to detect an effect of f = 0.86. This effect size we consider biologically relevant to see a relevant

reduction in immobility behaviour in the FST. The full R code for these calculations is provided in the appendix.

7.2.3 Prebiotics Administration

The prebiotic and control treatment was administered for 28 consecutive days (4 weeks). The treatments were administered within the first hour after lights off, the first hour of the animals' active phase. We tested the commercially available prebiotic product "Bimuno®" Powder (Bimuno, United Kingdom), which contains Galactooligosaccharides (B-GOS). A dose of 4 g/kg dissolved in tap water was administered to each animal per day, via syringe feeding. The dose was adjusted each week according to the weight of the animals. This prebiotic was chosen due to its superior effect on behaviour over FOS (Savignac et al., 2013). This dose was given to recreate the findings in previous literature (Savignac et al., 2013; Williams et al., 2016).

7.2.4 Control Administration

The control for the prebiotics was a mixture consisting of similar components to the Bimuno powder but without the active ingredient B-GOS. The control consisted of 50% lactose, 27% glucose, and 23% galactose (purchased from Sigma Aldrich). The control was administered at a dose of 4 g/kg/day, following the dosing regimen of previous literature (Williams et al., 2016; Savignac et al., 2016). The control was administered at the same time as the prebiotics, via syringe feeding, with the dose adjusted each week according to the weight of the animals. Animals were weighed each week, the weight was average per cage, and cages were divided into two groups, higher and lower weighing cage. The dose of prebiotic was calculated based on higher and lower cage weight averages. The table (7.1) below reports the doses given to each animal each week.

Table 7.1 Doses given to each animal each week.

Week	Lower Dose	Higher Dose
Week 1	1.25g	1.46g
Week 2	1.29g	1.55g
Week 3	1.37g	1.61g
Week 4	1.36g	1.67g

7.2.5 Syringe-feeding Details

Treatment was administered via syringe-feeding. The prebiotic, within a sweetened vehicle of glucose, was mixed with tap water, and added to a syringe. This is a newly established method for the accurate individual dosing of probiotics in rats (Tillmann & Wegener, 2017). With a training phase of roughly 3–4 days, to allow the rats to become accustomed to the administration and the taste, the rats willingly consume the mixture and approach the edge of the cage when the syringe is presented. This new method has been used for volumes of probiotic + vehicle solution up to 3 ml. This method of administration has been chosen to reduce the stress associated with oral gavage, and to increase the accuracy of dosing with administration of prebiotics in drinking water. In this experiment, the prebiotic Bimuno was added to tap water to give a total volume of 1.5 ml, as the smaller the volume, the sweeter the solution, which is thought to be more desirable for the rats to consume. Animals were fed at the start of the active cycle, time point +0 hours.

7.2.6 Measures to Reduce the Risk of Bias

7.2.6.1 Randomisation & Allocation Concealment

On the first day of the experiment, animals were moved from the breeding facility into the experimental facility. Animals were pair-housed; the two animals in each cage were the same strain and received the same treatment. Cages were randomly assigned to a group, Treatment or Control, to ensure allocation concealment during the handling and administration of treatment throughout the experiment. Randomisation was carried out using block randomisation with the online tool, The

Sealed Envelope (https://www.sealedenvelope.com/simple-randomiser/v1/lists), by a colleague not involved in the day-to-day running of the experiment. Cages were labelled with a unique randomisation code (e.g. GU9, LI3, etc) and a colour which signified the treatment given. This was done to minimise potential unconscious bias by 'remembering' which cages get which colour of treatment. Treatments were identified as red or blue. The cages (the experimental unit) were assigned randomly to treatment and the observational unit was the individual animal where the outcome of interest is measured. The observational unit (the animal) is nested within the experimental unit. The order of the cages was randomised in the racks at the beginning of the experiment to reduce possible effects from air-conditioning vents and/or being closer to the door. The placement of the cage was not be taken into consideration as a variable during analysis of the outcome data. Animals had 7 days to acclimatise to new housing facilities. The experimenter was blinded to which solution (prebiotic or vehicle control) each rat receives, as powders for the whole experiment were transferred to identical tub containers and labelled as 'blue' or 'red' by a colleague not involved in the experiment.

7.2.6.2 Blinded Assessment of Primary Outcome

The primary outcome was the FST. This outcome is recorded on video and scored manually. The videos were assessed blinded, before the group identity of the animals is revealed. The open field test and the elevated plus maze were analysed using the automated Ethovision scoring. All videos were analysed after all behavioural outcomes had been carried out. The primary experimenter was formally unblinded to the true group identity after the output of the data analysis files had been uploaded to the Open Science Framework (OSF) project.

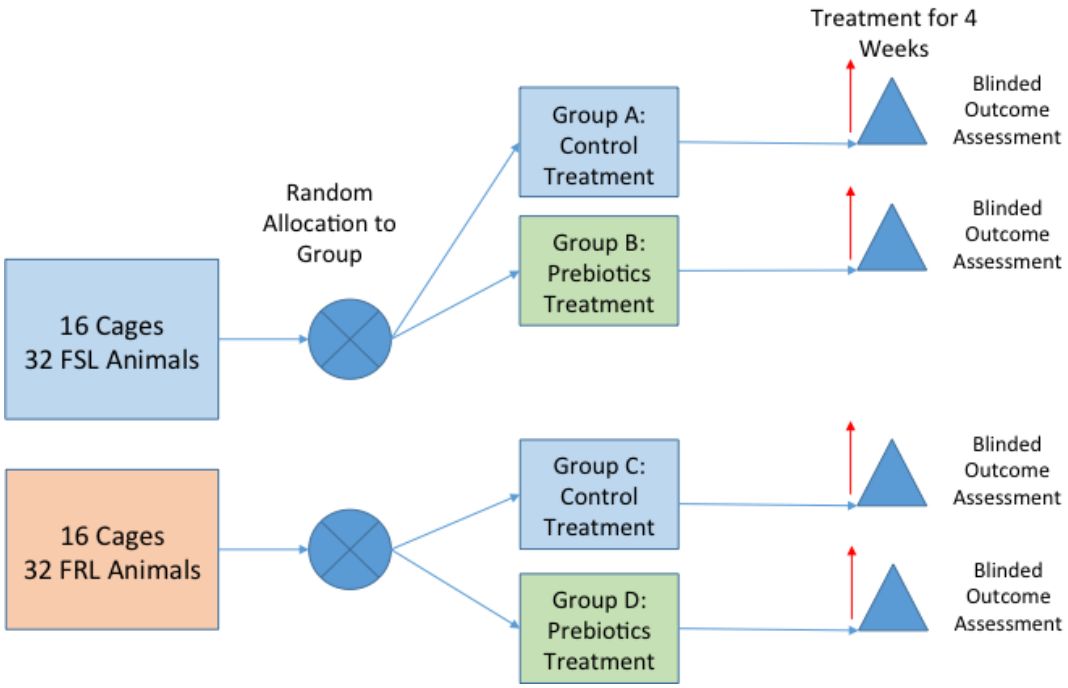


Figure 7.1. Experimental Design Setup

7.2.7 Outcome Assessment:

Behavioural assessment occurred during the rats' active phase, starting approximately 1 hour after administration of prebiotics, time point +1 hour, and lasting approximately 3 hours, until time point +4 hours.

7.2.7.1 Forced Swim Test

The primary outcome was performance on the forced swim test. On day 26 of the experiment, at time point +1 hr, at the start of the animals' active phase, the FST was be performed. The set-up consisted of 4 clear glass cylinders (60 cm in height × 24cm in diameter), separated with black opaque dividing walls, filled with water up to 40 cm. The temperature of the water was maintained at 25 °C ± 1°C. On the first day, the pre-swim session, the animals were placed in the tanks for 15 minutes. On the second day of testing, animals were placed into the tanks for 5 minutes. Both sessions were recorded by video camera. Both testing sessions were conducted in red light conditions, the rooms were only illuminated with red light at a wavelength of 625–740 nm as rats are less sensitive to red light (Peirson et al., 2018). The water was changed

in-between each test. Three behavioural parameters were assessed from the video footage, passive behaviour, immobility, and 2 active behaviours, swimming and climbing behaviour. Passive behaviour is defined as “the rat making no further movements beyond those needed to keep its head above the water” (Abildgaard et al., 2017). For each 5 second period, the predominant behaviour was recorded (immobility, swimming, or climbing). All FST sessions were scored by an experimenter blinded to the group assignment of the animals.

7.2.7.2 Open Field Test

Locomotor activity in the open field test (OFT) was assessed on day 27, immediately prior to the second FST session. Locomotor activity was assessed in a 100 cm × 100 cm (x 20 cm in height) black open field arena. Each animal was placed in the arena in the same starting location. Animals were assessed for 5 minutes in red light. All sessions were video-recorded and analysed using Noldus Ethovision XT9. Locomotor activity was assessed as the distance each animal moved in centimetres. The arena was cleaned with ethanol between each animal. All video recordings were scored by an experimenter blinded to the group assignment of the animals using the Noldus Ethovision automated scoring.

7.2.7.3 Elevated Plus Maze

Anxiety behaviour was assessed on day 24 in the elevated plus maze. The plus-shaped maze has two open arms and two closed arms (length: 50 cm x width: 10cm) and the centre zone measures 10 cm x 10 cm. Each animal was placed in the arena in the centre, facing the same open arm. Animals were assessed for 5 minutes and were tested during their active phase at time point +1hr. The light intensity in the open arms was 80–100 lx and 20 lx in the closed arms. Animals were kept in an adjacent dark experimental room and moved individually into the bright experimental room for testing. All sessions were video-recorded and analysed using Noldus Ethovision XT9. Anxiety behaviour was measured by calculating the time spent in the open arms in proportion to the time spent in the open arms and closed arms;

($Proportion = \frac{time\ spent\ in\ open\ arms}{time\ spent\ in\ open\ arms\ +\ time\ spent\ in\ closed\ arms}$); and number of entries into the open arms (defined as entire body of rat in the open arm). The arena was

cleaned with ethanol between each animal. All video recordings were scored by an experimenter blinded to the group assignment of the animals using the Noldus Ethovision automated scoring.

7.2.7.4 Body Weight & Food Consumption

Animals were weighed every week throughout the experiment, to assess if prebiotics administration influences weight gain, and also to adjust the dose of the prebiotics or control administered (4g/kg). Weekly food and water intake in the home cage was recorded.

7.2.7.5 Microbiota Analysis

Fecal boli were collected at the start of the experiment (day 1) and on the day of euthanisation (day 28). Fecal boli were collected directly from the anus of each animal into sterile tubes on dry ice, and frozen and stored at −80˚C. Fecal boli were intended to be analysed using 16S RNA sequencing to assess the composition of gut microbiota. These analyses were not carried out as there was an error during the shipping process and samples were identified without dry-ice in the package. As there may have been differential thawing of the sample and differential growth of bacterial communities within the fecal matter, we determined that the findings from the analyses would not be reliable and were therefore not carried out.

7.2.8 Exclusion Criteria

Exclusion criteria were pre-specified. Animals would be excluded if they displayed illness behaviour as assessed by trained, in house veterinarians. Animals would be excluded from statistical analysis if there were technical difficulties with video recording equipment or video files had been corrupted. No animals would be excluded from the statistical analysis if they successfully complete all aspects of the study. One animal was found dead in the home cage on day 10 of the experiment.

7.2.9 Experimental Procedure

Step 1: Animals were bred in house. Rats were moved into experimental housing and had a 7-day acclimatisation/habitation period to the animal housing facility prior to the start of the experiment. Cages were randomised into 2 groups; Treatment & Control.

Step 2: On day 1 of the experiment, each rat was weighed and fecal boli were collected. Then the first administration of prebiotics or control was given.

Step 3: Animals remained continuously on this treatment regimen for 4 weeks (28 days). Rats were weighed every week. The animals' daily food and water consumption was recorded.

Step 4: On day 24 animals were subjected to the elevated plus maze.

Step 5: On day 26 animals were subjected to the pre-swim of the FST.

Step 6: On day 27 the open field test and the FST were carried out.

Step 7: On day 28 animals were euthanized.

Tabl 7.2 The timing of outcome measure administration to each group.

Model	**Treatment**	**Length of Admin**	**Elevated Plus Maze**	**FST & Open Field**	**Euthanasia**
FSL	Prebiotics	4 weeks	Day 24	Day 27	Day 28
FSL	Control	4 weeks	Day 24	Day 27	Day 28
FRL	Prebiotics	4 weeks	Day 24	Day 27	Day 28
FRL	Control	4 weeks	Day 24	Day 27	Day 28

7.2.10 Data Analysis Pipeline

Data from the open field and FST were analysed in Ethovison. Videos were stored on an internal network drive. Video from each animal were scored blinded to the animal's group assignment. Data cleaning and statistical analysis was carried out in R studio (R, v.3.4.3).

7.2.11 Statistical Analysis

To test the hypothesis that probiotics improve depressive-like behaviour on the FST and anxiety-like behaviour in the EPM, the primary outcome being immobility time, a two-way ANOVA was conducted with 2 independent variables, treatment (control or prebiotic) and strain (FSL or FRL). Data were analysed at the level of experimental unit, the cage. Values from individual animals in the cage were averaged to give a cage value. Data were tested to see if they met the assumptions of two-way ANOVA. Levene's test was used to test homogeneity of variances. Shaprio-Wilk test was used to test normality of residuals. Data met the assumptions so a two-way ANOVA without any corrections was performed. All raw data and data analysis code were uploaded to OSF.

7.2.12 Amendments to Methods from Pre-specified Protocol

The protocol specified that 5-7 weeks old animals would be used. In this experiment animals between 8-14 weeks were used instead, as the breeding rate of the animals was not high enough to have 64 animals within a 2-week age bracket. There was a mistake in the protocol, which specified that the control would be the Bimuno Free Powder with the wrong composition, which is not a product. The control given was made up of similar components to the Bimuno powder but without the active ingredient B-GOS, consisting of 50% lactose, 27% glucose, and 23% galactose. The volume of the solution fed via syringe was reduced from 2ml to 1.5ml. The smaller the volume, the easier for the animals to consume, especially during the training of syringe feeding. I tested the weight of the powders each week to ensure that they properly dissolved in the amount of tap water, so that dosing was as accurate as possible. I administered the OFT for 5 minutes instead of 15mins specified in the protocol. The primary outcome of the OFT was to measure locomotor activity, which can be accurately assessed with a 5 minute testing period, so the length of the test was reduced to reduce the unnecessary stress to the animals. As there was no funding for neurochemical brain analyses, I did not dissect brains from the animals at euthanisation. I collected fecal boli on the first day behavioural testing (day 24) instead of on the day of euthanisation (day 28) with the aim of isolating the effects of prebiotics on gut microbiota, without the stress of the behavioural testing.

Further, it was not specified in the protocol that correction for multiple testing would be applied. With 7 outcomes investigated, the conventional level of significance being alpha = 0.05, when a Holm-Bonferroni correction to account for multiple testing is applied, the level of significance is 0.007.

7.3 Results

I was interested in the effects of prebiotics on seven outcomes; immobile, swimming, and struggling behaviour in the forced swim test, total distance moved in the open field test, proportion of entries to the open arms and frequency of whole body entries to the open arms in the elevated plus maze, and weight gain across the experiment.

7.3.1 Forced Swim Test

The primary outcome was immobility in the forced swim test as a measure of depressive-like behaviour in response to an inescapable stressful situation. There was no significant difference in time spent immobile between the FSL animals and the FRL animals ($F(1,28) = 8.09$, $p = 0.008$) at the corrected alpha level. There was no effect of the prebiotics on immobility ($F(1,28) = 3.34$, $p = 0.07$). Struggling behaviour differed significantly between the FSL animals and the FRL animals ($F(1,28) = 18.05$, $p = 0.0002$), with FRL animals showing more struggling behaviour than FSL animals. There was no effect of the prebiotics ($F(1,28) = 0.32$, $p = 0.57$). No significant differences between model or drug groups were observed in swimming behaviour.

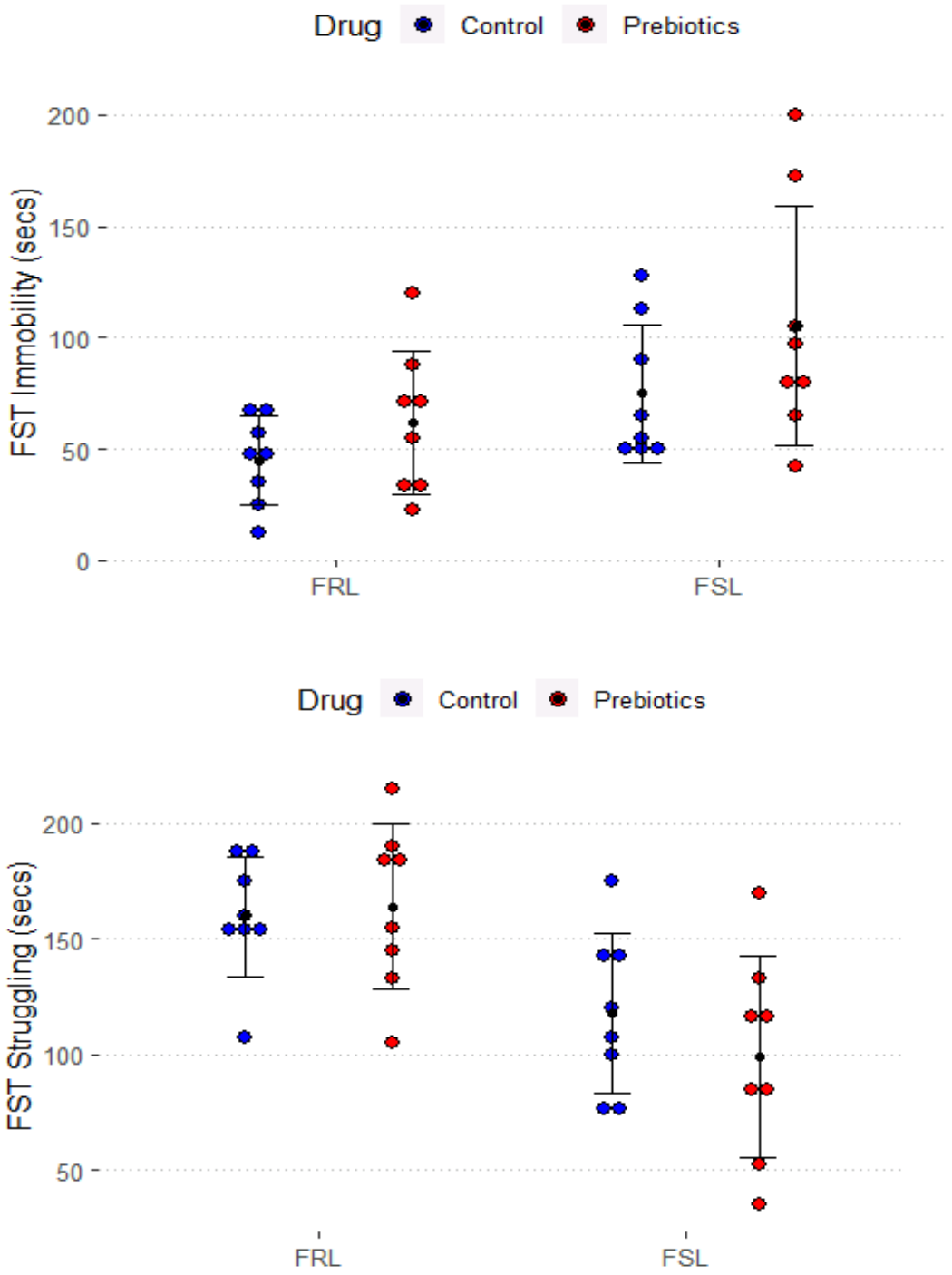


Figure 7.2A&B. Mean time spent A) Immobile (top) and B) Struggling (bottom) in the Forced Swim Test. Red (prebiotics) and blue (control) dots indicate performance of individual cages, black dots denote the mean for each group, error bars = SD, n = 8 experimental units per group.

7.3.2 Open Field Test

Locomotor activity was measured by total distance travelled in the open field test. There was a significant difference in the total distance travelled between the FSL animals and the FRL animals ($F(1,28) = 29.89$, $p < 0.0001$), FSL animals travelled further than the FRL animals. There was no effect of the prebiotics ($F(1,28) = 0.003$, $p = 0.95$).

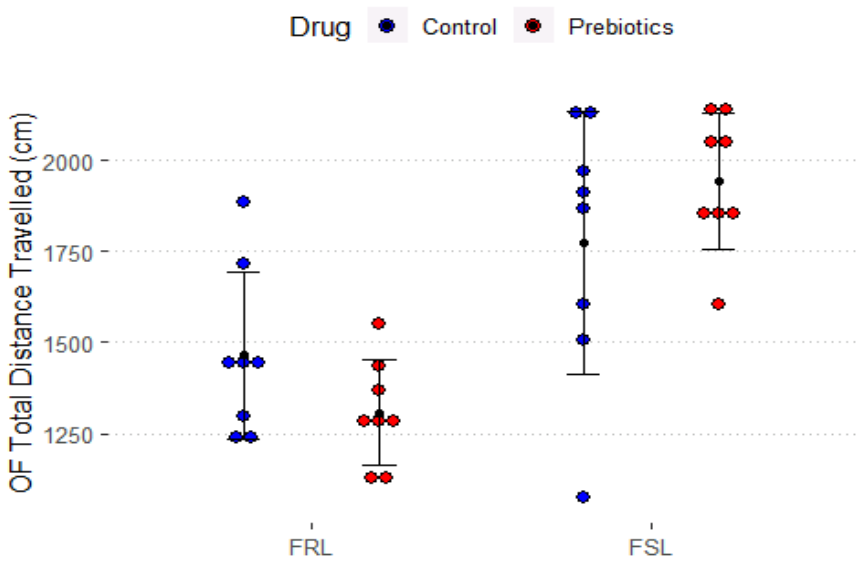


Figure 7.3. Mean distance travelled in the *Open Field test (5mins). Red (prebiotics) and blue (control) dots indicate performance of individual cages, black dots denote the mean for each group, error bars = SD, n = 8 experimental units per group.*

7.3.3 Elevated Plus Maze

There were no significant differences between model or drug groups in the proportion of time spent in the open arms. There was a significant difference in FSL and FRL animals in the number of entries into open arms, defined as entire body of rat in the open arm, ($F(1,28) = 8.52$, $p = 0.006$), with no effects of treatment ($F(1,28) = 0.03$, $p = 0.86$).

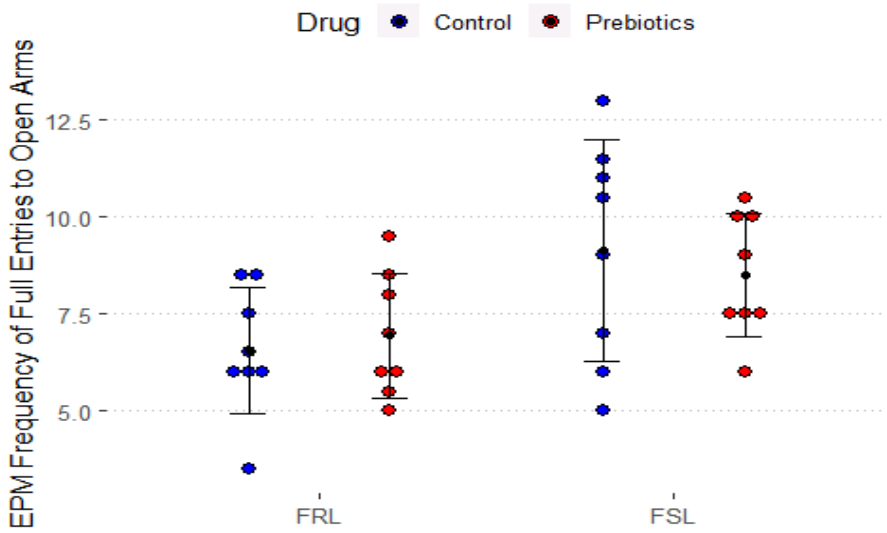


Figure 7.4. Mean frequency of full body entries to the open arms in the Elevated Plus Maze. *Red (prebiotics) and blue (control) dots indicate performance of individual cages, black dots denote the mean for each group, error bars = SD, n = 8 experimental units per group.*

7.3.4 Weight

At the beginning of the experiment, all the groups had different weights. The mean weight of the FRL Prebiotics group was highest, mean weight in FSL Prebiotics and FRL Control were roughly similar, and the mean weight of the FSL Control group was lowest (Table 7.3). I calculated the weight gain across the experiment (weight at euthanisation minus the starting weight). There was a significant difference in FRL and FSL animals in weight gain ($F(1,28) = 37.57$, $p < 0.0001$). There were no significant differences between treatment groups or significant interactions between treatment and model.

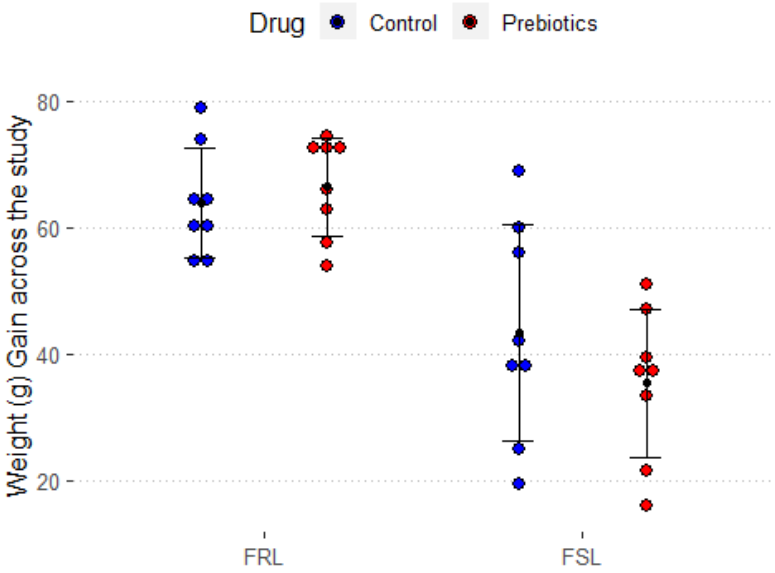


Fig 7.5. Mean weight (g) gain for each group across the study. Red (prebiotics) and blue (control) dots indicate performance of individual cages, black dots denote the mean for each group, error bars = SD, n = 8 experimental units per group.

Table 3. Mean and SD values for all outcomes.

Outcome	FSL Control	FSL Prebiotics	FRL Control	FRL Prebiotics
FST Immobility (secs)	75.0 (31.11)	105.31 (53.99)	45.0 (19.77)	61.87 (32.51)
FST Swimming (secs)	107.5 (24.64)	95.63 (29.15)	95.31 (23.24)	74.38 (16.73)
FST Struggling (secs)	117.5 (34.62)	99.0 (43.79)	159.69 (25.75)	163.75 (35.81)
OF total distance (cm)	1774.62 (360.85)	1941.77 (186.50)	1464.22 (230.26)	1307.12 (144.87)
EPM Proportion time spent in open arms (%)	23.89 (10.58)	31.13 (19.29)	31.32 (13.87)	32.35 (13.25)
EPM Frequency of full body entries to open arms (freq)	9.13 (2.86)	8.50 (1.60)	6.56 (1.64)	6.94 (1.59)
Weight (g) Week 1	317.50 (20.53)	329.19 (24.62)	333.38 (24.92)	374.94 (33.29)
Weight (g) Week 2	333.31 (20.19)	342.63 (23.25)	363.69 (24.51)	408.19 (28.97)
Weight (g) Week 3	349.44 (21.79)	360.31 (24.91)	386.50 (26.30)	431.06 (31.46)
Weight (g) Week 4	351.00 (19.96)	362.63 (24.13)	391.25 (26.39)	437.19 (29.83)
Weight (g) Euthanisation	360.94 (20.19)	364.56 (23.19)	397.38 (26.79)	441.44 (30.01)

All values displayed are mean (SD).

7.3.5 Post-hoc Analysis

I decided, post-hoc, to analyse the data for the primary outcomes, behaviour on the forced swim test, omitting correction for multiple testing and using the typical unit of analysis, the individual animal.

When no correction for multiple testing was applied, there was a significant difference in time spent immobile between the FSL animals and the FRL animals ($F(1,28) = 8.09$, $p = 0.008$), FSL animals spent more time immobile than FRL animals. There was no effect of the prebiotics on immobility ($F(1,28) = 3.34$, $p = 0.07$). Struggling behaviour differed significantly between the FSL animals and the FRL animals ($F(1,28) = 18.05$, $p = 0.0002$), with FRL animals showing more struggling behaviour than FSL animals. There was no effect of the prebiotics ($F(1,28) = 0.32$, $p = 0.57$). No significant differences between model or drug groups were observed in swimming behaviour.

When data for the primary outcome, behaviour on the forced swim test, were analysed with the individual animal as the experimental unit and without correction for multiple testing, the effect of treatment on immobility in the forced swim test was significant, ($F(1,59) = 5.37$, $p = 0.02$), with animals receiving prebiotics spending more time immobile in both FSL and FRL animals. There was a significant difference between treatment groups in swimming behaviour ($F(1,59) = 4.22$, $p = 0.045$), with animals receiving prebiotics spending less time swimming than animals receiving control. There was a significant difference between model groups in struggling behaviour ($F(1,59) = 23.97$, $p < 0.0001$), with FSL animals spending less time struggling than FRL animals. There was no significant effect of treatment on struggling behaviour. All of the significant effects of treatment, apart from model differences in struggling behaviour, are not significant with correction for multiple testing.

7.4 Discussion

I administered B-GOS prebiotics or a glucose vehicle substance to FSL and FRL animals for 4 weeks. I observed behavioural differences between the FSL and FRL groups; as expected, the control Flinders Resistant Line animals displayed more struggling behaviour in the FST, as a measure of active escape behaviour. I found that FSL animals travelled further in the open field test, displaying hyperlocomotion. FSL animals displayed more full-body entries to the open arms, as a proxy for anxiety-like behaviour. As there was a difference between the model groups in locomotor behaviour in the open field test, the difference in groups in the frequency of entries to the open arms is likely an artefact of the increased locomotion in the FSL group. When

the proportion of time spent in the open arms was calculated, there were no significant differences between the model or treatment groups.

This study employed measures to reduce the risk of bias, including random allocation of animals to treatment or control groups, allocation concealment throughout the experiment and the treatment administration and behavioural testing, the outcomes were assessed blinded, and an *a priori* sample size calculation was performed to calculate the required power and number of animals required to detect an effect in this experiment. The experimental unit, the smallest unit the treatment could independently be given, in this study was the cage. Interestingly, the post-hoc analysis carried out where analysis was performed on the individual animal level data, treatment had a significant effect. The statistical methods used to analyse the data assume independence of samples (Altman & Bland, 1997). Incorrect analysis of data that are not independent leads to increased false positive rates (type 1 error; Parsons et al., 2017).

Information from a systematic review and meta-analysis (Chapter 5) informed this study. As of 2016 there was a range of probiotics investigated, but no prebiotics were tested, and the most commonly used behavioural outcome measure used was the forced swim test.

I did not observe any significant effects of the B-GOS prebiotics in any depressive-like or anxiety-like outcomes. I did see an effect of the prebiotics in differentially affecting weight at the model groups at the time of euthanisation, with animals receiving prebiotics weighing more than the control groups in the FRL animals but not in the FSL model of depression.

All four groups started off the experiment at significantly different weights. In general, it is observed in the breeding colonies that FRL animals weigh more than FSL animals (Overstreet et al., 2005). There were already differences between drug groups, with prebiotic groups weighing more than vehicle groups. Investigating differences in weight at euthanisation, there appears to be a differential effect of the treatment

administration, with prebiotics increasing the weight in the FRL group but with no differences between groups observed in the FSL group. The differential weight differences are after 28 days of administration of prebiotics. A follow-up study could investigate the effects of the prebiotics after 8 weeks of administration.

There are various possible reasons why I did not see effects of the prebiotics in this study. Gut microbiota differ between species, the relative abundances of most of the dominant genera are for example quite different in mice and humans (Nyugen et al., 2015). Rat gut microbiota (Sprague Dawley, the ancestral strain of the FSL and FRL) has been shown to have 46 bacterial species, with relative abundance of the bacterial communities progressively decreasing with age (Flemer et al., 2017). It may be that the prebiotic has little effect in rodents due to differences between humans and rats in gut microbiota composition. Prebiotics and probiotics are being investigated in randomised controlled trials in depressed patients and a recent systematic review and meta-analysis concluded that they are overall effective (Huang et al., 2016). Further, the age range of the animals used in this study (8-14 weeks old) may contribute to the large variability seen in behaviour, and therefore the lack of significant effects of the prebiotics, as age can change the composition of the bacterial communities in the gut (Flemer et al., 2017). A similar study using probiotics in FSL and FRL animals found no effects of probiotics administered alone but when animals were on a high fat diet, FSL animals displayed reduced immobility behaviour in the forced swim test (Abildgaard et al., 2017b). B-GOS stimulate the growth of *Bifidobacterium.* It is possible that at the start of the experiment, animals did not have high numbers of this colony of bacteria, and therefore the relative benefit of B-GOS is small.

The forced swim test was the outcome measure chosen to assess depressive-like behaviour in this experiment as it is the most widely used in the field. It may however, not be the most effective measure. The FST measures the response to a test, an inescapable situation, which may reflect learning or coping and may not be reflective of a long-term depressive state in the animals (Molendijk & de Kloet, 2015; Commons et al., 2017). Future studies utilising more naturalistic outcomes such as social behaviour in the home cage may improve our understanding of depressive-like behaviour in animals.

A key limitation of this study is the lack of biological outcome assessments. The secondary outcome measure of this study was to identify the impact of B-GOS administration on gut microbiota beta diversity and to observe any changes in microbial communities before and after administration of the prebiotics. Without biological readouts, I am unable to investigate whether the lack of behavioural effects of the prebiotics were accompanied by similar effects in gut microbiota diversity. Without analysis of microbiota composition, I am unable to ascertain whether there were initial differences in the FSL and FRL animals that may have contributed to differential effects of the prebiotic, which no study has done to date. Future studies can investigate the effects of prebiotics not only on gut microbiota changes, but also on inflammatory pathways and neurotransmitter levels. Inflammatory and immune system responses have been observed in animal models of depression that are altered through gut-microbiota targeting interventions (Abildgaard et al., 2017a, Abildgaard et al., 2017b). The investigation into the effects of prebiotics on inflammatory and immune responses would add to this literature.

It was not pre-specified in the protocol that correction for multiple testing should be taken into account at the analysis stage. In addition, the sample size calculation was carried out not taking into account the nested structure of the experimental design, where an individual animal was nested or clustered in a cage and treatments were assigned at cage level. The sample size calculation should have been corrected with a variance inflation factor to account for the correlation between observations within clusters (Parsons et al., 2017). The variance inflation factor relies on knowing the intra-cluster correlation coefficient from previous data or an estimate (Parsons et al., 2017). This study could have been underpowered to detect effects due to inaccurate a priori sample size calculation. I did not control for litter effects in the random allocation of animals to treatment groups. Individual animals or cages may not have been fully independent, which is another factor that could have confounded the results from statistical analysis.

A further limitation is that the control and the prebiotics were largely composed of sugars, which are likely involved in neural mechanisms in reward, therefore, the

behavioural measures are not independent of the effects of sugar. Daily sugar bingeing has been shown to increase dopamine in nucleus acccumbens (Rada et al., 2005). The nucleus accumbens is a key area involved in the reward system and anti-depressant mechanisms of drugs (Nestler, 2015). Future studies could employ an additional control group of saline treated or naïve animals, to understand the impact of the sugars in the prebiotics and their effect on depressive-like behaviours.

7.5 Conclusion

In conclusion, I investigated the effects of 4-week administration of B-GOS prebiotics or a glucose vehicle substance in FSL and FRL animals on depressive-like and anxiety-like behaviour. No effects of the prebiotics were observed. Follow-up studies can be conducted to investigate the prolonged effects of prebiotics on depressive-like behaviour in FSL animals. Prebiotics are being investigated in randomised controlled trials with depressed patients and may prove effective if no effects are seen in rodent studies as there are significant differences in the composition of gut microbiota between rodents and humans (Flemer et al., 2017).

8 General Discussion & Conclusions

The aim of this thesis was to use systematic review and meta-analysis techniques to synthesise the evidence available on animal models of depression and antidepressants investigated, to achieve an overview of what species are selected, and what techniques are used to induce depressive-like phenotypes.

I conducted this research because despite the decades of research into depression, current treatments are adequate at best and the underlying pathological mechanisms are not yet understood. Depression remains a leading source of disability with an estimated 322 million people worldwide are suffering with depression (WHO, 2017).

A systematic search of PubMed and EMBASE in May 2016 identified 70,365 unique publications that were potentially relevant to animal models of depression (Chapter 2). The field of animal models of depression was larger than initially anticipated, and it was not feasible to manually conduct a systematic review of this literature within the time-frame of a PhD, let alone the first step of title and abstract screening. Therefore, I developed, tested, and implemented automation tools, with the help of collaborators, to the screening stage (Chapter 3) and the annotation stage (Chapter 4) of this systematic review of animal models of depression. The machine learning tools performed well and were implemented successfully at the screening stage. The machine learning approaches were trained on 5,749 records with human screening decisions. The best performing algorithm achieved a performance of 98.7% sensitivity and 88.3% specificity. The machine learning approach screened 63,365 records that did not have human inclusion decisions, reducing the human time required for this task by an estimated 40 person months. Using this machine learning approach, I was interested to test if machine learning could help identify human screening errors. The machine learning algorithm was applied to the records with human decisions using k-fold validation. Errors were identified and corrected, which when retraining the algorithm using the corrected human decisions, improving performance of the algorithms. This is a novel approach in the context of systematic review methodology. With further testing, this approach may be implemented routinely in the existing methodological framework of systematic review methods.

Machine learning greatly assisted in reducing the human resources required to screen publications for inclusion, identifying 18,409 publications that were highly likely to be relevant to animal models of depression. This is still a large amount of unique citations to manually extract information from. Therefore, I developed automatic annotation tools to assist in annotating and grouping these documents by the key terms, method of disease model induction, and antidepressant treatments. Dictionaries of key terms were converted into regular expression dictionaries and the dictionaries were applied to the title, abstracts, and where available, full-text PDFs of the 18,409 included studies. This approach was used in three systematic reviews to identify relevant studies for further systematic review and meta-analysis e.g. ketamine (Chapter 6). This approach may be limited in that without full-text PDFs key outcomes or interventions may not be identified. The accuracy of these dictionaries is being further investigated in ongoing reviews. Despite these limitations, this approach has proved beneficial and highlights the need for domain-agreed ontologies to help categorise the increasing amount of literature in animal models of neuropsychiatric disease.

Utilising the automation tools developed, this thesis presents findings from two systematic reviews, the first on the available literature on microbiota-targeting interventions in animal models of depression (Chapter 5), and the second on the effects of ketamine as an antidepressant in literature on animal models of depression (Chapter 6). The review of microbiota-targeting interventions identified few investigations with a broad range of outcomes and interventions, with no two studies investigating the same probiotic strain. This broad range of data likely reflects the recent interest in investigating the role of gut microbiota in depression, and the exploration of potential mechanisms involved in the effects of interventions, such as probiotics, on depressive-like outcomes. The reporting of measures to reduce the risk of bias in these studies was relatively high in comparison with previous reviews of animal models of other neurological diseases. The studies mainly reported using male animals, limiting the generalisability of the findings in primary studies. There was not enough data to fully analyse the impact of timing of administration. The investigation into whether probiotics may promote resilience to stressful conditions or can reduce depressive-like outcomes after exposure to stressful events, may prove useful for treatment in humans. This is a topic for investigation with further animal studies.

The findings from the systematic review of microbiota-targeting interventions highlighted gaps in the literature where additional animal studies can contribute, the investigation of prebiotics had not been identified in the review and existing studies were at risk of bias. The findings from the systematic review were used to inform the design of a primary animal experiment investigating the effects of prebiotics in Flinders Sensitive Line animals (Chapter 7). The effect of the galacto-oligosaccharide prebiotics, administered for 28 days, on depressive-like and anxiety-like behaviour was investigated in a genetic model of depression. An *a priori* sample size calculation, blinding, and randomisation were employed to reduce the risk of bias. No effect of prebiotics was seen on behaviour in depressive-like rats, although differences in weight were seen between groups. A follow-up study with longer administration of prebiotics, and with gut microbiota and neurochemical outcome assessment may further elucidate the impact of prebiotics in depressive-like outcomes.

The systematic review of ketamine as an antidepressant in literature on animal models of depression reported data from 182 experiments investigating the effects of ketamine on behaviour in the forced swim test. Data showed an overall antidepressant effect of ketamine and an increase in timing between ketamine administration and outcome assessment was associated with significant decreases in effect size. Future animal studies in ketamine can improve the value of animal experiments by exploring aspects of study design that may impact effect in clinical trial such as, the timing of administration, the route of administration, sex. There was statistically significant heterogeneity, which may have been due to the imprecision of the primary studies included in the review. With this wide variation seen, a way forward might be to establish multi-centre trials of animal models of neuropsychiatric disorders, ensuring that any significant effects observed are robust to controlled across-laboratory variation.

These reviews analysed and discussed the internal and external validity of animal models of depression and antidepressant treatments tested. These findings can be used to better understand factors that impact the efficacy of treatments in animal models and improve the quality of research conducted, to better understand

depression and improve the treatments available to patients. Findings from both the systematic review of microbiota targeting interventions and the systematic review of ketamine clearly highlight the poor quality of reporting of measures to reduce the risk of bias in studies describing animal models of depression. These findings call for the conduct of higher quality animal studies. Although the field of depression shows slightly higher levels of reporting of measures to reduce the risk of bias than the existing body of literature describing animal studies of neurological disorders, this is by no means adequate. I hope that the continuing efforts from several stakeholders including researchers, journals, and funders will ensure that animal research conducted and published adheres to reporting guidelines such as the ARRIVE guidelines. This will help to improve the quality of studies on animal models of depression. However, evidence from studies investigating the implemention of reporting guidelines for animal studies in top journals (Hair et al., 2018; Macleod, 2017) shows that mandating the completion of checklists does little to improve actual reporting. Therefore varied approaches are needed to for a sustainable improvement in reporting. As academic culture shifts and open science practices such as pre-registration of animal experiments become more prevalent, meta-analysis will be able to summarise more accurately the true effect size of an intervention and not rely on statistical techniques such as Egger's regression to adjust for negative and neutral unpublished studies.

The systematic reviews in this PhD have highlighted concerns of external validity in animal models of depression. It is common in animal experiments describing anti-depressant interventions intended for clinical use that interventions are administered prior to the model induction. For treatments such as probiotics that have fewer recorded adverse side effects this may be feasible in a clinical setting. However, for interventions such as ketamine where severe adverse side-effects have been recorded and patients must be monitored intensively when this treatment is administered through an IV, this is less feasible. The clinical trials that have commenced to test the administration of intra-nasal ketamine have also noted patients experiencing adverse side-effects. Research in animal models can help inform clinical trial design by systematically exploring variables to ensure that external validity of the experiment is increased. Systematically investigating variables such as dose of administration, route of administration, and the sex of the animals can improve

external validity. In addition to measuring the primary outcome of effect of an antidepressant, also measuring outcomes that assess potential side-effects in humans may ensure that significant findings may be more easily translated to a clinical setting.

The application of automation tools to this systematic review of animal models of depression has allowed for two systematic reviews of the field to be conducted since 2016. Although these tools are under development, their utility has been highlighted. This dataset created using these tools is freely accessible online (Chapter 4). This allows other researchers to make use of the dataset to carry out further reviews of animal models of depression. The datasets made freely available as part of the application of machine learning algorithms for citation screening (Chapter 3) can be used as a validation set by developers building new tools. With further refinement and with the implementation of living systematic reviews, this dataset can be continuously updated with the latest research in the field of animal models of depression. The implementation of living systematic reviews is a goal of the International Collaboration for the Automation of Systematic Reviews (ICASR) and several research groups are working towards this aim. Specifically, the next stage in automating systematic review methods is being carried out by the SLIM collaboration to reduce the human effort required to extract data from graphs in primary articles (Crammod et al., 2018). Data extraction from primary articles is carried out in duplicate as human error in this stage can lead to false conclusions being drawn about the data (Mathes et al., 2017). Machine-assisted data extraction from graphs may reduce this error and reduce the amount of resources required to carry out this step in the systematic review process.

In future, I hope that with the development of novel automation tools and refinement of existing tools, systematic review and meta-analysis methodology can be applied to the entire knowledge base of animal models of depression. This can only be achieved when the human resources required to perform this methodology are sufficiently reduced. Pooling data from the model induction techniques and from antidepressant treatments investigated will allow us to gain a full summary of the depression in animals and further understand the mechanisms behind neuropsychiatric disease. The application of novel automation tools can be applied to improve animal modelling in other fields. An example of this is the multi-centre work being carried out from

collaborative approaches to generate shared ontologies and dictionaries (Nielson et al., 2015; Callahan et al., 2016). With key terms mapped across an entire knowledge base, this allows for the investigation of network connections and mapping within a large dataset, including the correlations between behavioural and biological outcomes. This can be used to generate new biological hypotheses. Pooling data from the studies investigating antidepressants in animal models of depression could allow for the application of advanced meta-analysis techniques such as network meta-analysis. This tool applied to clinical systematic review data can allow comparisons to be made between the effects in drugs when direction comparisons have not been carried out in primary studies (Cipriani et al., 2018). This technique may be useful in the context of animal models of depression.

In future, the cultural changes associated with open science, data sharing, and automation tools can enable high quality, high speed evidence synthesis in the age of information explosion. Olkin's remark in 1995, that meta-analysis is the key to dealing with the increasing amounts of literature, has never been so pertinent. As a field, we must work harder to overcome the barriers to understanding complex disorders such as depression.

REFERENCES

Abildgaard, A., Elfving, B., Hokland, M., Wegener, G. and Lund, S., 2017. Probiotic treatment reduces depressive-like behaviour in rats independently of diet. *Psychoneuroendocrinology*, *79*, pp.40-48. (A)

Abildgaard, A., Elfving, B., Hokland, M., Lund, S. and Wegener, G., 2017. Probiotic treatment protects against the pro-depressant-like effect of high-fat diet in Flinders Sensitive Line rats. *Brain, behavior, and immunity*, *65*, pp.33-42. (B)

Abrams, R., 2002. *Electroconvulsive therapy*. Oxford University Press.

Abramson, L.Y., Metalsky, G.I. and Alloy, L.B., 1989. Hopelessness depression: A theory-based subtype of depression. *Psychological review*, *96*(2), p.358.

Albert, P. R., Benkelfat, C., & Descarries, L. (2012). The neurobiology of depression—revisiting the serotonin hypothesis. I. Cellular and molecular mechanisms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1601), 2378–2381. http://doi.org/10.1098/rstb.2012.0190

Altman, D.G. and Bland, M., 1997. Statistics notes: units of analysis. *Bmj*, *314*(7098), p.1874.

American Psychiatric Association. 2013. Diagnostic and statistical manual of mental disorders (5th ed.). Arlington, VA: American Psychiatric Publishing.

Anacker, C., Zunszain, P.A., Cattaneo, A., Carvalho, L.A., Garabedian, M.J., Thuret, S., Price, J. and Pariante, C.M., 2011. Antidepressants increase human hippocampal neurogenesis by activating the glucocorticoid receptor. *Molecular psychiatry*, *16*(7), p.738.

Autry, A.E., Adachi, M., Nosyreva, E., Na, E.S., Los, M.F., Cheng, P.F., Kavalali, E.T. and Monteggia, L.M., 2011. NMDA receptor blockade at rest triggers rapid behavioural antidepressant responses. *Nature*, *475*(7354), p.91.

Bahor, Z. , Nunes-Fonseca, C. , Currie, G. L. , Sena, E. S. , Thomson, L. D. , Macleod, M. R., 2015. 'Improving Our Understanding of txhe in vivo Modelling of Psychotic Disorders'. *World Academy of Science, Engineering and Technology, International Science Index, Medical and Health Sciences,* 9(7), p. 619.

Bahor, Z., Nunes-Fonseca, C., Thomson, L.D., Sena, E.S. and Macleod, M.R., 2016. Improving our understanding of the in vivo modelling of psychotic disorders: A protocol for a systematic review and meta-analysis. *Evidence-based Preclinical Medicine*, *3*(2), pp.10-16.

Bahor, Z., Liao, J., Macleod, M.R., Bannach-Brown, A., McCann, S.K., Wever, K.E., Thomas, J., Ottavi, T., Howells, D.W., Rice, A. and Ananiadou, S., 2017. Risk of bias reporting in the recent animal focal cerebral ischaemia literature. *Clinical Science*, *131*(20), pp.2525-2532.

Bannach-Brown, A., Liao, J., Wegener, G., Macleod, M. R., 2016. Understanding in vivo modelling of depression: a systematic review protocol. *CAMARADES repository of protocols. Retrieved from:* https://drive.google.com/file/d/0BxckMffc78BYLWM2QUpBY3l1Q1k/view

Bannach-Brown, A., Thomas, J., Przybyła, P., Liao, J., (2016). "Protocol for Error Analysis: Machine learning and text mining solutions for systematic reviews of animal

models of depression". Published on CAMARADES Website. www.CAMARADES.info. Direct Access: https://drive.google.com/file/d/0BxckMffc78BYTm0tUzJJZkc1alk/view

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A.S., Ananiadou, S., Liao, J. and Macleod, M.R., 2018. The use of text-mining and machine learning algorithms in systematic reviews: reducing workload in preclinical biomedical sciences and reducing human screening error. *bioRxiv*, p.255760.

Belzung, C., 2014. Innovative drugs to treat depression: did animal models fail to be predictive or did clinical trials fail to detect effects?. *Neuropsychopharmacology*, *39*(5), p.1041.

Belzung, C. and Lemoine, M., 2011. Criteria of validity for animal models of psychiatric disorders: focus on anxiety disorders and depression. *Biology of mood & anxiety disorders*, *1*(1), p.9.

Belzung, C., Willner, P. and Philippot, P., 2015. Depression: from psychopathology to pathophysiology. *Current opinion in neurobiology*, *30*, pp.24-30.

Berman, R.M., Cappiello, A., Anand, A., Oren, D.A., Heninger, G.R., Charney, D.S. and Krystal, J.H., 2000. Antidepressant effects of ketamine in depressed patients. *Biological psychiatry*, *47*(4), pp.351-354.

Benca, R.M., Overstreet, D.E., Gilliland, M.A., Russell, D., Bergmann, B.M. and Obermeyer, W.H., 1996. Increased basal REM sleep but no difference in dark induction or light suppression of REM sleep in Flinders rats with cholinergic supersensitivity. *Neuropsychopharmacology*, *15*(1), p.45.

Bimuno®, 2018. United Kindgom. https://www.bimuno.com/ Accessed 20th April 2018

Biondi-Zoccai, G., Lotrionte, M., Landoni, G. and Modena, M.G., 2011. The rough guide to systematic reviews and meta-analyses. *HSR proceedings in intensive care & cardiovascular anesthesia*, *3*(3), p.161.

Blanchard, R.J. and Blanchard, D.C., 1977. Aggressive behavior in the rat. *Behavioral biology*, 21(2). pp.197-224

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), pp.993-1022.

Borah, R., Brown, A.W., Capers, P.L. and Kaiser, K.A., 2017. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ open*, *7*(2), p.e012545.

Borenstein, M., Hedges, L.V., Higgins, J.P. and Rothstein, H.R., 2011. *Introduction to meta-analysis*. John Wiley & Sons.

Bornmann, L. and Mutz, R., 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), pp.2215-2222.

Bravo, J.A., Forsythe, P., Chew, M.V., Escaravage, E., Savignac, H.M., Dinan, T.G., Bienenstock, J. and Cryan, J.F., 2011. Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. *Proceedings of the National Academy of Sciences*, p.201102999.

Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

Bunney, W.E. and Bunney, B.G., 2000. Molecular clock genes in man and lower animals: possible implications for circadian abnormalities in depression. *Neuropsychopharmacology*, *22*(4), p.335-45.

Burokas, A., Arboleya, S., Moloney, R.D., Peterson, V.L., Murphy, K., Clarke, G., Stanton, C., Dinan, T.G. and Cryan, J.F., 2017. Targeting the microbiota-gut-brain axis: prebiotics have anxiolytic and antidepressant-like effects and reverse the impact of chronic stress in mice. *Biological psychiatry*, *82*(7), pp.472-487.

Caldarone, B. and Brunner, D., 2009. Preclinical testing in mice for treatment resistant depression. *Front Neurosci*, *3*, pp.264-265.

Caldarone, B.J., Zachariou, V. and King, S.L., 2015. Rodent models of treatment-resistant depression. *European journal of pharmacology*, *753*, pp.51-65.

Callahan, A., Abeyruwan, S.W., Al-Ali, H., Sakurai, K., Ferguson, A.R., Popovich, P.G., Shah, N.H., Visser, U., Bixby, J.L. and Lemmon, V.P., 2016. RegenBase: a knowledge base of spinal cord injury biology for translational research. *Database*, *2016*.

Callahan, A., Anderson, K.D., Beattie, M.S., Bixby, J.L., Ferguson, A.R., Fouad, K., Jakeman, L.B., Nielson, J.L., Popovich, P.G., Schwab, J.M. and Lemmon, V.P., 2017. Developing a data sharing community for spinal cord injury research. *Experimental neurology*, *295*, pp.135-143.

Carney, S., Cowen, P., Geddes, J., Goodwin, G., Rogers, R., Dearness, K., Tomlin, A., Eastaugh, J., Freemantle, N., Lester, H. and Harvey, A., 2003. Efficacy and safety of electroconvulsive therapy in depressive disorders: a systematic review and meta-analysis. *The Lancet*. 361, pp. 799–808

Carter, A.R., Tilling, K. and Munafò, M.R., 2017. A systematic review of sample size and power in leading neuroscience journals. *bioRxiv*, p.217596.

Celie, J.E., Loeys, T., Desmet, M. and Verhaeghe, P., 2017. The Depression Conundrum and the Advantages of Uncertainty. *Frontiers in psychology*, *8*, p.939.

Chaudhury, D., Walsh, J.J., Friedman, A.K., Juarez, B., Ku, S.M., Koo, J.W., Ferguson, D., Tsai, H.C., Pomeranz, L., Christoffel, D.J. and Nectow, A.R., 2013. Rapid regulation of depression-related behaviours by control of midbrain dopamine neurons. *Nature*, *493*(7433), p.532.

Chen, F., Madsen, T.M., Wegener, G. and Nyengaard, J.R., 2010. Imipramine treatment increases the number of hippocampal synapses and neurons in a genetic animal model of depression. *Hippocampus*, *20*(12), pp.1376-1384.

Chen, J., Yang, C., Guo, B., Sena, E.S., Macleod, M.R., Yuan, Y. and Hirst, T.C., 2016. The efficacy of trastuzumab in animal models of breast cancer: A systematic review and meta-analysis. *PloS one*, *11*(7), p.e0158240.

Chesney, E., Goodwin, G.M. and Fazel, S., 2014. Risks of all-cause and suicide mortality in mental disorders: a meta-review. *World Psychiatry*, *13*(2), pp.153-160.

Chuang, J.C. and Zigman, J.M., 2010. Ghrelin's roles in stress, mood, and anxiety regulation. *International journal of peptides*, epub 14 Feb *2010*.

Cipriani, A., Furukawa, T.A., Salanti, G., Geddes, J.R., Higgins, J.P., Churchill, R., Watanabe, N., Nakagawa, A., Omori, I.M., McGuire, H. and Tansella, M., 2009.

Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *The lancet*, *373*(9665), pp.746-758.

Cipriani, A., Furukawa, T.A., Salanti, G., Chaimani, A., Atkinson, L.Z., Ogawa, Y., Leucht, S., Ruhe, H.G., Turner, E.H., Higgins, J.P. and Egger, M., 2018. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet*, *391*(10128), pp.1357-1366.

Cohen, A.M., Hersh, W.R., Peterson, K. and Yen, P.Y., 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, *13*(2), pp.206-219.

Cohen, A.M., Adams, C.E., Davis, J.M., Yu, C., Yu, P.S., Meng, W., Duggan, L., McDonagh, M. and Smalheiser, N.R., 2010, November. Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. In *Proceedings of the 1st ACM international Health Informatics Symposium* (pp. 376-380). ACM.

Cohen, A.M., Ambert, K. and McDonagh, M., 2012. Studying the potential impact of automated document classification on scheduling a systematic review update. *BMC medical informatics and decision making*, *12*(1), p.33.

Collobert, R. and Weston, J., 2008, July. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160-167). ACM.

Colman, I., Jones, P.B., Kuh, D., Weeks, M., Naicker, K., Richards, M. and Croudace, T.J., 2014. Early development, stress and depression across the life course: pathways to depression in a national British birth cohort. *Psychological medicine*, *44*(13), pp.2845-2854.

Commons, K.G., Cholanians, A.B., Babb, J.A. and Ehlinger, D.G., 2017. The rodent forced swim test measures stress-coping strategy, not depression-like behavior. *ACS chemical neuroscience*, *8*(5), pp.955-960.

Comroe, J.H. and Dripps, R.D., 1976. Scientific basis for the support of biomedical science. *Science, 192*(4235), pp.105-111.

ContentMine Tools, 2016. Accessed from: https://contentmine.github.io/ on 05/09/2016.

Contopoulos-Ioannidis, D.G., Ntzani, E. and Ioannidis, J.P., 2003. Translation of highly promising basic science research into clinical applications. *The American journal of medicine*, *114*(6), pp.477-484.

Conway, C.R., George, M.S. and Sackeim, H.A., 2017. Toward an evidence-based, operational definition of treatment-resistant depression: when enough is enough. *JAMA psychiatry*, *74*(1), pp.9-10.

Cormack, G.V. and Grossman, M.R., 2016, July. Engineering quality and reliability in technology-assisted review. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 75-84). ACM.

Cramond, F., O'Mara-Eves, A., Doran-Constant, L., Rice, A.S., Macleod, M. and Thomas, J., 2018. The development and evaluation of an online application to assist

in the extraction of data from graphs for use in systematic reviews. *Wellcome Open Research*, *3*, p. 157. (https://doi.org/10.12688/wellcomeopenres.14738.1)

Cryan, J.F. and Mombereau, C., 2004. In search of a depressed mouse: utility of models for studying depression-related behavior in genetically modified mice. *Molecular psychiatry*, *9*(4), p.326.

Cryan, J.F. and O'Mahony, S.M., 2011. The microbiome-gut-brain axis: from bowel to behavior. *Neurogastroenterology & Motility*, *23*(3), pp.187-192.

Cryan, J.F. and Slattery, D.A., 2007. Animal models of mood disorders: recent developments. *Current opinion in psychiatry*, *20*(1), pp.1-7.

Currie, G.L., Delaney, A., Bennett, M.I., Dickenson, A.H., Egan, K.J., Vesterinen, H.M., Sena, E.S., Macleod, M.R., Colvin, L.A. and Fallon, M.T., 2013. Animal models of bone cancer pain: systematic review and meta-analyses. *PAIN*®, *154*(6), pp.917-926.

Deakin, J.W., Lees, J., McKie, S., Hallak, J.E., Williams, S.R. and Dursun, S.M., 2008. Glutamate and the neural basis of the subjective effects of ketamine: a pharmaco–magnetic resonance imaging study. *Archives of general psychiatry*, *65*(2), pp.154-164.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R., 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), pp.391-407.

de Vries, R.B., Hooijmans, C.R., Tillema, A., Leenaars, M. and Ritskes-Hoitinga, M., 2011. A search filter for increasing the retrieval of animal studies in Embase. *Laboratory animals*, *45*(4), pp.268-270.

de Vries, R.B., Wever, K.E., Avey, M.T., Stephens, M.L., Sena, E.S. and Leenaars, M., 2014. The usefulness of systematic reviews of animal experiments for the design of preclinical and clinical studies. *ILAR journal*, *55*(3), pp.427-437.

de Vries, R.B., Hooijmans, C.R., Tillema, A., Leenaars, M. and Ritskes-Hoitinga, M., 2014. Letter to the Editor: Updated version of the Embase search filter for animal studies, Laboratory Animals, 48(1).

de Vries, R.B., Hooijmans, C.R., Langendam, M.W., van Luijk, J., Leenaars, M., Ritskes-Hoitinga, M. and Wever, K.E., 2015. A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies. *Evidence-based Preclinical Medicine*, *2*(1), pp.1-9.

Dinan, T.G., Stanton, C. and Cryan, J.F., 2013. Psychobiotics: a novel class of psychotropic. *Biological psychiatry*, *74*(10), pp.720-726.

Dinan, T.G. and Cryan, J.F., 2017. The microbiome-gut-brain axis in health and disease. *Gastroenterology Clinics*, *46*(1), pp.77-89.

Duval, S. and Tweedie, R., 2000. Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), pp.455-463.

Dzirasa, K. and Covington, H.E., 2012. Increasing the validity of experimental models for depression. *Annals of the New York Academy of Sciences*, *1265*(1), pp.36-45.

Eastman, C.I., Young, M.A., Fogg, L.F., Liu, L. and Meaden, P.M., 1998. Bright light treatment of winter depression: a placebo-controlled trial. *Archives of general psychiatry*, *55*(10), pp.883-889.

Egan, K.J., Vesterinen, H.M., Beglopoulos, V., Sena, E.S. and Macleod, M.R., 2016. From a mouse: systematic analysis reveals limitations of experiments testing interventions in Alzheimer's disease mouse models. *Evidence-based Preclinical Medicine*, *3*(1), pp.12-23.

Egger, M., Smith, G.D., Schneider, M. and Minder, C., 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ*, *315*(7109), pp.629-634.

EndNote. Version X7.2. & XT11 Thomson-Reuters. www.endnote.com

Eyre, H. and Baune, B.T., 2012. Neuroplastic changes in depression: a role for the immune system. *Psychoneuroendocrinology*, *37*(9), pp.1397-1416.

Feldman, K.A., 1971. Using the work of others: Some observations on reviewing and integrating. *Sociology of Education*, *44,* pp.86-102.

Flather, M., 2015. Open access data sharing from clinical trials: is it really feasible?. *European Heart Journal - Quality of Care and Clinical Outcomes, 1(2),* pp. 49-50

Flemer, B., Gaci, N., Borrel, G., Sanderson, I.R., Chaudhary, P.P., Tottey, W., O'Toole, P.W. and Brugère, J.F., 2017. Fecal microbiota variation across the lifespan of the healthy laboratory rat. *Gut microbes*, *8*(5), pp.428-439.

Francino, M.P., 2016. Antibiotics and the human gut microbiome: dysbioses and accumulation of resistances. *Frontiers in microbiology*, *6*, p.1543.

Fried, E.I., 2017. Moving forward: how depression heterogeneity hinders progress in treatment and research. Expert Review of Neurotherapeutics, 17(5), pp. 423-425. DOI: 10.1080/14737175.2017.1307737

Fried, E.I., 2015. Problematic assumptions have slowed down depression research: why symptoms, not syndromes are the way forward. *Frontiers in psychology*, *6*, p.309.

Fried, E.I. and Nesse, R.M., 2015. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC medicine*, *13*(1), p.72. (A)

Fried, E.I. and Nesse, R.M., 2015. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR* D study. *Journal of affective disorders*, *172*, pp.96-102. (B)

Fried, E.I., Nesse, R.M., Zivin, K., Guille, C. and Sen, S., 2014. Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors. *Psychological medicine*, *44*(10), pp.2067-2076.

Friedman, J., Hastie, T. and Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, *33*(1), p.1.

Gartlehner, G., Gaynes, B.N., Amick, H.R., Asher, G.N., Morgan, L.C., Coker-Schwimmer, E., Forneris, C., Boland, E., Lux, L.J., Gaylord, S. and Bann, C., 2016. Comparative benefits and harms of antidepressant, psychological, complementary, and exercise treatments for major depression: an evidence report for a clinical practice guideline from the American College of Physicians. *Annals of internal medicine*, *164*(5), pp.331-341.

Geddes, J.R., Carney, S.M., Davies, C., Furukawa, T.A., Kupfer, D.J., Frank, E. and Goodwin, G.M., 2003. Relapse prevention with antidepressant drug treatment in depressive disorders: a systematic review. *The Lancet*, *361*(9358), pp.653-661.

Geyer MA, Markou A (1995). Animal models of psychiatric disorders. In: Psychopharmacology: fourth generation of progress. Bloom FE, Kupfer D, (editors). New York: Raven, pp. 787–798. http://www.acnp.org/g4/GN401000076/Default.htm

Gibson, G.R., Hutkins, R., Sanders, M.E., Prescott, S.L., Reimer, R.A., Salminen, S.J., Scott, K., Stanton, C., Swanson, K.S., Cani, P.D. and Verbeke, K., 2017. Expert consensus document: The International Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of prebiotics. *Nature Reviews Gastroenterology and Hepatology*, *14*(8), p.491.

Gillman, P.K., 2007. Tricyclic antidepressant pharmacology and therapeutic drug interactions updated. *British journal of pharmacology*, *151*(6), pp.737-748.

Green, S.M. and Coté, C.J., 2009. Ketamine and neurotoxicity: clinical perspectives and implications for emergency medicine. *Annals of emergency medicine*, *54*(2), pp.181-190.

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E., Dodel, R., Ekman, M., Faravelli, C., Fratiglioni, L. and Gannon, B., 2011. Cost of disorders of the brain in Europe 2010. *European neuropsychopharmacology*, *21*(10), pp.718-779.

Hair, K., Macleod, M.R., Sena, E.S., & The IICARus Collaboration. "A randomised controlled trial of an Intervention to Improve Compliance with the ARRIVE guidelines (IICARus)" bioRxiv 370874; doi: https://doi.org/10.1101/370874

Hara, K. and Matsumoto, Y., 2007. Extracting clinical trial design information from MEDLINE abstracts. *New Generation Computing*, *25*(3), pp.263-275.

Harald, B. and Gordon, P., 2012. Meta-review of depressive subtyping models. *Journal of affective disorders*, *139*(2), pp.126-140.

Hamon, M. and Blier, P., 2013. Monoamine neurocircuitry in depression and strategies for new treatments. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *45*, pp.54-63.

Hanson, N.D., Owens, M.J. and Nemeroff, C.B., 2011. Depression, antidepressants, and neurogenesis: a critical reappraisal. *Neuropsychopharmacology*, *36*(13), p.2589.

Harnad, S., 2007. Ethics of open access to biomedical research: just a special case of ethics of open access to research. *Philosophy, Ethics, and Humanities in Medicine*, *2*(1), p.31.

Hedges, L.V., Tipton, E. and Johnson, M.C., 2010. Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods*, *1*(1), pp.39-65.

Henkel, V., Mergl, R., Allgaier, A.K., Kohnen, R., Möller, H.J. and Hegerl, U., 2006. Treatment of depression with atypical features: a meta-analytic approach. *Psychiatry research*, *141*(1), pp.89-101.

Henn, F.A. and Vollmayr, B., 2005. Stress models of depression: forming genetically vulnerable strains. *Neuroscience & Biobehavioral Reviews*, *29*(4-5), pp.799-804.

Hersh, J.K., 2013. Electroconvulsive therapy (ECT) from the patient's perspective. *Journal of medical ethics*, *39*(3), pp.171-172.

Higgins, J.P., Thompson, S.G., Deeks, J.J. and Altman, D.G., 2003. Measuring inconsistency in meta-analyses. *BMJ: British Medical Journal*, *327*(7414), p.557.

Hillhouse, T.M. and Porter, J.H., 2015. A brief history of the development of antidepressant drugs: From monoamines to glutamate. *Experimental and clinical psychopharmacology*, *23*(1), p.1.

Holsboer, F., Von Bardeleben, U., Wiedemann, K., Müller, O.A. and Stalla, G.K., 1987. Serial assessment of corticotropin-releasing hormone response after dexamethasone in depression implications for pathophysiology of DST nonsuppression. *Biological psychiatry*, *22*(2), pp.228-234.

Hooijmans, C.R., Tillema, A., Leenaars, M. and Ritskes-Hoitinga, M., 2010. Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. *Laboratory animals*, *44*(3), pp.170-175.

Hooijmans, C.R., Rovers, M.M., de Vries, R.B., Leenaars, M., Ritskes-Hoitinga, M. and Langendam, M.W., 2014. SYRCLE's risk of bias tool for animal studies. *BMC medical research methodology*, *14*(1), p.43.

Horn, J., De Haan, R.J., Vermeulen, M., Luiten, P.G.M. and Limburg, M., 2001. Nimodipine in animal model experiments of focal cerebral ischemia: a systematic review. *Stroke*, *32*(10), pp.2433-2438.

Howard, B.E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M.R., Holmgren, S., Pelch, K.E., Walker, V., Rooney, A.A. and Macleod, M., 2016. SWIFT-Review: a text-mining workbench for systematic review. *Systematic reviews*, *5*(1), p.87.

Hsu, W., Speier, W. and Taira, R.K., 2012. Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature. In *AMIA Annual Symposium Proceedings* (Vol. 2012, p. 350). American Medical Informatics Association.

Huang, R., Wang, K. and Hu, J., 2016. Effect of probiotics on depression: a systematic review and meta-analysis of randomized controlled trials. *Nutrients*, *8*(8), p.483.

Huynh, N.N. and McIntyre, R.S., 2008. What are the implications of the STAR* D trial for primary care? A review and synthesis. *Primary care companion to the Journal of clinical psychiatry*, *10*(2), p.91.

Insel, T.R., 2006. Beyond efficacy: the STAR* D trial. *American Journal of Psychiatry*, *163*(1), pp.5-7.

Insel, T., 2013. Post by former NIMH director Thomas Insel: Transforming diagnosis. *National Institute of Mental Health*. Retrieved from: https://www.nimh.nih.gov/about/directors/thomas-insel/blog/2013/transforming-diagnosis.shtml

Ioannidis, J.P., 2009. Ranking antidepressants. *The Lancet*, *373*(9677), pp.1759-1760.

Ioannidis, J.P., 2012. Extrapolating from animals to humans. *Science Translational Medicine*, *4*(151), pp.151.

Iwai, T., Ohnuki, T., Sasaki-Hamada, S., Saitoh, A., Sugiyama, A. and Oka, J.I., 2013. Glucagon-like peptide-2 but not imipramine exhibits antidepressant-like effects in ACTH-treated mice. *Behavioural brain research*, *243*, pp.153-157.

Janssen PAJ. In: Animal behavior and drug action. A Ciba Foundation Symposium. Steinberg H, de Reuck AVS, Knight J, editor. London: J.A. Churchill Ltd; 1964. Screening tests and prediction from animals to man; pp. 264–268.

Jonnalagadda, S.R., Goyal, P. and Huffman, M.D., 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic reviews*, *4*(1), p.78.

Jue, T.R., Sena, E.S., Macleod, M.R., McDonald, K.L. and Hirst, T.C., 2018. A systematic review and meta-analysis of topoisomerase inhibition in pre-clinical glioma models. *Oncotarget*, *9*(13), p.11387.

Kara, N., Stukalin, Y. and Einat, H., 2017. Revisiting the validity of the mouse forced swim test: systematic review and meta-analysis of the effects of prototypic antidepressants. *Neuroscience & Biobehavioral Reviews*.

Kashani, P., Yousefian, S., Amini, A., Heidari, K., Younesian, S. and Hatamabadi, H.R., 2014. The effect of intravenous ketamine in suicidal ideation of emergency department patients. *Emergency*, *2*(1), p.36.

Katz, R.J., Roth, K.A. and Carroll, B.J., 1981. Acute and chronic stress effects on open field activity in the rat: implications for a model of depression. *Neuroscience & Biobehavioral Reviews*, *5*(2), pp.247-251.

Kelly, C. and Yang, H., 2013. A system for extracting study design parameters from nutritional genomics abstracts. *Journal of integrative bioinformatics*, *10*(2), pp.82-93.

Kennedy, R.J., Kirk, S.J. and Gardiner, K.R., 2001. Probiotics (Br J Surg 2001; 88: 161-2). *The British journal of surgery*, *88*(7), pp.1018-1019.

Kerr, K.F., Wang, Z., Janes, H., McClelland, R.L., Psaty, B.M. and Pepe, M.S., 2014. Net reclassification indices for evaluating risk-prediction instruments: a critical review. *Epidemiology (Cambridge, Mass.)*, *25*(1), p.114.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K.R., Rush, A.J., Walters, E.E. and Wang, P.S., 2003. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA*, *289*(23), pp.3095-3105.

Kessler, R.C., Berglund, P., Demler, O., Jin, R., Merikangas, K.R. and Walters, E.E., 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry*, *62*(6), pp.593-602.

Kieburtz, K. and Olanow, C.W., 2007. Translational experimental therapeutics: The translation of laboratory-based discovery into disease-related therapy. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine: A Journal of Translational and Personalized Medicine*, *74*(1), pp.7-14.

Kilkenny, C., Browne, W.J., Cuthill, I.C., Emerson, M. and Altman, D.G., 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS biology*, *8*(6), p.e1000412.

Kim, H.J., McGuire, D.B., Tulman, L. and Barsevick, A.M., 2005. Symptom clusters: concept analysis and clinical implications for cancer nursing. *Cancer nursing*, *28*(4), pp.270-282.

Kim, H., Bian, J., Mostafa, J., Jonnalagadda, S. and Del Fiol, G., 2016. Feasibility of extracting key elements from ClinicalTrials. gov to support clinicians' patient care decisions. In *AMIA Annual Symposium Proceedings* (Vol. 2016, p. 705). American Medical Informatics Association.

King, L.A., 2003. Behavioral evaluation of the psychological welfare and environmental requirements of agricultural research animals: theory, measurement, ethics, and practical implications. *ILAR journal*, *44*(3), pp.211-221.

Kiritchenko, S., de Bruijn, B., Carini, S., Martin, J. and Sim, I., 2010. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, *10*(1), p.56.

Kitamura, Y., Araki, H., Suemaru, K. and Gomita, Y., 2002. Effects of imipramine and lithium on wet-dog shakes mediated by the 5-HT2A receptor in ACTH-treated rats. *Pharmacology Biochemistry and Behavior*, *72*(1-2), pp.397-402.

Klarer, M., Arnold, M., Günther, L., Winter, C., Langhans, W. and Meyer, U., 2014. Gut vagal afferents differentially modulate innate anxiety and learned fear. *Journal of Neuroscience*, *34*(21), pp.7067-7076.

Koenigs, M. and Grafman, J., 2009. Posttraumatic stress disorder: the role of medial prefrontal cortex and amygdala. *The Neuroscientist*, *15*(5), pp.540-548.

Kontonatsios, G., Brockmeier, A.J., Przybyła, P., McNaught, J., Mu, T., Goulermas, J.Y. and Ananiadou, S., 2017. A semi-supervised approach using label propagation to support citation screening. *Journal of biomedical informatics*, *72*, pp.67-76.

Koob GF, Heinrichs SC, Britton K. In: The American Psychiatric Press Textbook of Psychopharmacology. Eds. Schatzberg, A.F. and Nemeroff, C.B., 1995. *The American Psychiatric Press textbook of psychopharmacology*. American Psychiatric Association.

Krishnan, K.R., 2007. Revisiting monoamine oxidase inhibitors. *The Journal of clinical psychiatry*, *68*, pp.35-41.

Krishnan, V. and Nestler, E.J., 2008. The molecular neurobiology of depression. *Nature*, *455*(7215), p.894-902.

Krishnan, V. and Nestler, E.J., 2011. Animal models of depression: molecular perspectives. In *Current Top* (pp. 121-147). Springer, Berlin, Heidelberg.

Krogh, J., Hjorthøj, C., Speyer, H., Gluud, C. and Nordentoft, M., 2017. Exercise for patients with major depression: a systematic review with meta-analysis and trial sequential analysis. *BMJ open*, *7*(9), p.e014820.

Kuhn, M., 2017. "The caret package". Retrieved from: https://topepo.github.io/caret/

Landhuis, E., 2016. Scientific literature: Information overload. *Nature*, *535*(7612), pp.457-458.

Landis, S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal, R.G., Darnell, R.B., Ferrante, R.J., Fillit, H. and Finkelstein, R., 2012. A call

for transparent reporting to optimize the predictive value of preclinical research. Nature, 490(7419), pp.187-191.

Lang, U.E. and Borgwardt, S., 2013. Molecular mechanisms of depression: perspectives on new treatment strategies. *Cellular Physiology and Biochemistry*, *31*(6), pp.761-777.

Lawson, E.A., Miller, K.K., Blum, J.I., Meenaghan, E., Misra, M., Eddy, K.T., Herzog, D.B. and Klibanski, A., 2012. Leptin levels are associated with decreased depressive symptoms in women across the weight spectrum, independent of body fat. *Clinical endocrinology*, *76*(4), pp.520-525.

Lee, B.H. and Kim, Y.K., 2010. The roles of BDNF in the pathophysiology of major depression and in antidepressant treatment. *Psychiatry investigation*, *7*(4), pp.231-235.

Leenaars, M., Hooijmans, C.R., van Veggel, N., Ter Riet, G., Leeflang, M., Hooft, L., van der Wilt, G.J., Tillema, A. and Ritskes-Hoitinga, M., 2012. A step-by-step guide to systematically identify all relevant animal studies. *Laboratory animals*, *46*(1), pp.24-31.

Lewis, D.D. and Gale, W.A., 1994, August. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 3-12). Springer-Verlag New York, Inc

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P., Clarke, M., Devereaux, P.J., Kleijnen, J. and Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS medicine*, *6*(7), p.e1000100.

Lichtenberg, P. and Belmaker, R.H., 2010. Subtyping major depressive disorder. *Psychotherapy and psychosomatics*, *79*(3), pp.131-135.

Lin, J., 2018. Preprints growth rate ten times higher than journal articles. CrossRef. Retrieved from: https://www.crossref.org/blog/preprints-growth-rate-ten-times-higher-than-journal-articles/ on 07-07-2018

Linde, K., Sigterman, K., Kriston, L., Rücker, G., Jamil, S., Meissner, K. and Schneider, A., 2015. Effectiveness of psychological treatments for depressive disorders in primary care: systematic review and meta-analysis. *The Annals of Family Medicine*, *13*(1), pp.56-68. (A)

Linde, K., Rücker, G., Sigterman, K., Jamil, S., Meissner, K., Schneider, A. and Kriston, L., 2015. Comparative effectiveness of psychological treatments for depressive disorders in primary care: network meta-analysis. *BMC family practice*, *16*(1), p.103. (B)

Loomer, H.P., Saunders, J.C. and Kline, N.S., 1957. A clinical and pharmacodynamic evaluation of iproniazid as a psychic energizer. *Psychiatric research reports*. 8, pp.129-41.

López-Muñoz, F. and Alamo, C., 2009. Monoaminergic neurotransmission: the history of the discovery of antidepressants from 1950s until today. *Current pharmaceutical design*, *15*(14), pp.1563-1586.

Liao, J., Ananiadou, S., Currie, G.L., Howard, B.E., Rice, A., Sena, E.S., Thomas, J., Varghese, A. and Macleod, M.R., 2018. Automation of citation screening in pre-clinical systematic reviews. *bioRxiv*, p.280131.

Liu, J., Timsina, P. and El-Gayar, O., 2018. A comparative analysis of semi-supervised learning: The case of article selection for medical systematic reviews. *Information Systems Frontiers*, *20*(2), pp.195-207.

Liu, W., Ge, T., Leng, Y., Pan, Z., Fan, J., Yang, W. and Cui, R., 2017. The role of neural plasticity in depression: from hippocampus to prefrontal cortex. *Neural plasticity*, *2017*.

Macleod, M.R., O'Collins, T., Howells, D.W. and Donnan, G.A., 2004. Pooling of animal experimental data reveals influence of study design and publication bias. *Stroke*, *35*(5), pp.1203-1208.

Macleod, M.R., McLean, A.L., Kyriakopoulou, A., Serghiou, S., de Wilde, A., Sherratt, N., Hirst, T., Hemblade, R., Bahor, Z., Nunes-Fonseca, C. and Potluru, A., 2015. Risk of bias in reports of in vivo research: a focus for improvement. *PLoS biology*, *13*(10), p.e1002273.

Macleod, M.R., 2017. Findings of a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution. *bioRxiv*, p.187245.

Madsen, T.M., Treschow, A., Bengzon, J., Bolwig, T.G., Lindvall, O. and Tingström, A., 2000. Increased neurogenesis in a model of electroconvulsive therapy. *Biological psychiatry*, *47*(12), pp.1043-1049.

Manning, C. D., Raghavan, P., & Schütze, H. 2008. Introduction to Information Retrieval, Cambridge University Press: USA.

Marcus, M., Yasamy, M. T., van Ommeren, M., Chisholm, D., & Saxena, S., 2012. Depression: A global public health concern. *World Health Organisation*. Retrieved from http://www.who.int/mental_health/management/depression/who_paper_depression_wfmh_2012.pdf on 5th November, 2015

Marshall, I.J., Kuiper, J. and Wallace, B.C., 2015. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, *23*(1), pp.193-201.

Mathes, T., Klaßen, P. and Pieper, D., 2017. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *BMC medical research methodology*, *17*(1), p.152.

McArthur, R. and Borsini, F., 2006. Animal models of depression in drug discovery: a historical perspective. *Pharmacology Biochemistry and Behavior*, *84*(3), pp.436-452.

McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit." Retrieved from: http://mallet.cs.umass.edu.

McCann, S.K., Irvine, C., Mead, G.E., Sena, E.S., Currie, G.L., Egan, K.E., Macleod, M.R. and Howells, D.W., 2014. Efficacy of antidepressants in animal models of ischemic stroke: a systematic review and meta-analysis. *Stroke*, *45*(10), pp.3055-3063.

McGirr, A., Berlim, M.T., Bond, D.J., Fleck, M.P., Yatham, L.N. and Lam, R.W., 2015. A systematic review and meta-analysis of randomized, double-blind, placebo-controlled trials of ketamine in the rapid treatment of major depressive episodes. *Psychological medicine*, *45*(4), pp.693-704.

McGorry, P.D., Hickie, I.B., Yung, A.R., Pantelis, C. and Jackson, H.J., 2006. Clinical staging of psychiatric disorders: a heuristic framework for choosing earlier, safer and more effective interventions. *Australian and New Zealand Journal of Psychiatry*, *40*(8), pp.616-622.

McKiernan, E.C., Bourne, P.E., Brown, C.T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B.A., Ram, K., Soderberg, C.K. and Spies, J.R., 2016. How open science helps researchers succeed, ELife 5. *See https://doi. org/10.7554/elife*, *16800*.

McVey Neufeld, K.A., O'Mahony, S.M., Hoban, A.E., Waworuntu, R.V., Berg, B.M., Dinan, T.G. and Cryan, J.F., 2017. Neurobehavioural effects of Lactobacillus rhamnosus GG alone and in combination with prebiotics polydextrose and galactooligosaccharide in male rats exposed to early-life stress. *Nutritional neuroscience*, pp.1-10.

Merkl, A., Heuser, I. and Bajbouj, M., 2009. Antidepressant electroconvulsive therapy: mechanism of action, recent advances and limitations. *Experimental neurology*, *219*(1), pp.20-26.

Mertsalov, K., & McCreary, M. 2009. Document classification with support vector machines. *Rational Enterprise: White Paper*. Accessed from: http://www.rationalenterprise.com/assets/content/files/Classification_with_Support_Vector_Machines.pdf , on: 05/09/2016.

Mika, A., Day, H.E., Martinez, A., Rumian, N.L., Greenwood, B.N., Chichlowski, M., Berg, B.M. and Fleshner, M., 2017. Early life diets with prebiotics and bioactive milk fractions attenuate the impact of stress on learned helplessness behaviours and alter gene expression within neural circuits important for stress resistance. *European journal of neuroscience*, *45*(3), pp.342-357.

Millard, L.A., Flach, P.A. and Higgins, J.P., 2015. Machine learning to assist risk-of-bias assessments in systematic reviews. *International journal of epidemiology*, *45*(1), pp.266-277.

Michie, S., Thomas, J., Johnston, M., Mac Aonghusa, P., Shawe-Taylor, J., Kelly, M.P., Deleris, L.A., Finnerty, A.N., Marques, M.M., Norris, E. and O'Mara-Eves, A., 2017. The Human Behaviour-Change Project: harnessing the power of artificial intelligence and machine learning for evidence synthesis and interpretation. *Implementation Science*, *12*(1), p.121.

Miwa, M., Thomas, J., O'Mara-Eves, A. and Ananiadou, S., 2014. Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics*, *51*, pp.242-253.

Molendijk, M.L. and de Kloet, E.R., 2015. Immobility in the forced swim test is adaptive and does not reflect depression. *Psychoneuroendocrinology*, *62*, pp.389-391.

Mongeau, R., Hamon, M. and Lanfumey, L., 2011. How can stress alter emotional balance through its interaction with the serotonergic system?. *The handbook of stress: neuropsychological effects on the brain*, pp.480-504.

Monroe, S.M. and Simons, A.D., 1991. Diathesis-stress theories in the context of life stress research: implications for the depressive disorders. *Psychological bulletin*, *110*(3), p.406.

Montgomery, S.A. and Kasper, S., 1995. Comparison of compliance between serotonin reuptake inhibitors and tricyclic antidepressants: a meta-analysis. *International Clinical Psychopharmacology*.9 Suppl 4, pp.33-40.

MQ: Transforming Mental Health, 2016. Depression: Asking the Right Questions. Retrieved from: http://b.3cdn.net/joinmq/10fbab8ed26626f32e_a3m6bjx2e.pdf 22/01/2016 on 10/07/2018

Muller, J., Pentyala, S., Dilger, J. and Pentyala, S., 2016. Ketamine enantiomers in the rapid and sustained antidepressant effects. *Therapeutic advances in psychopharmacology*, *6*(3), pp.185-192.

Murray, C.J., Abraham, J., Ali, M.K., Alvarado, M., Atkinson, C., Baddour, L.M., Bartels, D.H., Benjamin, E.J., Bhalla, K., Birbeck, G. and Bolliger, I., 2013. The state of US health, 1990-2010: burden of diseases, injuries, and risk factors. Jama, 310(6), pp.591-606.

Murrough, J.W., 2012. Ketamine as a novel antidepressant: from synapse to behavior. *Clinical Pharmacology & Therapeutics*, *91*(2), pp.303-309.

Naci, H., Cooper, J. and Mossialos, E., 2015. Timely publication and sharing of trial data: opportunities and challenges for comparative effectiveness research in cardiovascular disease. *European Heart Journal–Quality of Care and Clinical Outcomes*, *1*(2), pp.58-65.

Nakazawa, M., 2018. “fsmb” Package. Retrieved from: https://cran.r-project.org/web/packages/fmsb/fmsb.pdf

National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs), 2018. *Experimental Unit*. https://eda.nc3rs.org.uk/experimental-design-unit, Accessed on: 03-05-2018.

National Health Service, 2014. Treating Clinical Depression. Retrieved from: http://www.nhs.uk/Conditions/Depression/Pages/Treatment.aspx on 05/09/2016

National Health Service, 2016. Treatments for Depression. Retrieved from: https://www.nhs.uk/conditions/clinical-depression/treatment/ on 10/07/2018

National Institute for Health and Clinical Excellence, 2009. Depression in Adults: Recognition and management. London: National Institute for Health and Clinical Excellence. Retrieved from: https://www.nice.org.uk/guidance/cg90 on 10/07/2018

National Collaborating Centre for Mental Health (UK), 2010. Depression: The Treatment and Management of Depression in Adults. Leicester, UK: British Psychological Society; APPENDIX 11, THE CLASSIFICATION OF DEPRESSION AND DEPRESSION RATING SCALES/QUESTIONNAIRES. Retrieved from: https://www.ncbi.nlm.nih.gov/books/NBK63740/ on 10/07/2018

Nguyen, T.L.A., Vieira-Silva, S., Liston, A. and Raes, J., 2015. How informative is the mouse for human gut microbiota research?. *Disease models & mechanisms*, *8*(1), pp.1-16.

Nielson, J.L., Paquette, J., Liu, A.W., Guandique, C.F., Tovar, C.A., Inoue, T., Irvine, K.A., Gensel, J.C., Kloke, J., Petrossian, T.C. and Lum, P.Y., 2015. Topological data

analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature communications*, *6*, p.8581.

Nestler, E.J., 2015. Role of the brain's reward circuitry in depression: transcriptional mechanisms. In *International review of neurobiology* (Vol. 124, pp. 151-170). Academic Press.

Nestler, E.J., Gould, E. and Manji, H., 2002. Preclinical models: status of basic research in depression. *Biological psychiatry*, *52*(6), pp.503-528.

Newcombe, R.G., 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in medicine*, *17*(8), pp.857-872.

Oades, R.D. and Halliday, G.M., 1987. Ventral tegmental (A10) system: neurobiology. 1. Anatomy and connectivity. *Brain Research Reviews*, *12*(2), pp.117-165.

O'Collins, V.E., Macleod, M.R., Donnan, G.A., Horky, L.L., van der Worp, B.H. and Howells, D.W., 2006. 1,026 experimental treatments in acute stroke. *Annals of neurology*, *59*(3), pp.467-477.

Olkin, I., 1995. Statistical and theoretical considerations in meta-analysis. *Journal of Clinical Epidemiology*, *48*(1), pp.133-146.

O'Mahony, S.M., Felice, V.D., Nally, K., Savignac, H.M., Claesson, M.J., Scully, P., Woznicki, J., Hyland, N.P., Shanahan, F., Quigley, E.M. and Marchesi, J.R., 2014. Disturbance of the gut microbiota in early-life selectively affects visceral pain in adulthood without impacting cognitive or anxiety-related behaviors in male rats. *Neuroscience*, *277*, pp.885-901.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M. and Ananiadou, S., 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, *4*(1), p.5.

Oracle, 2018. MySQL 8.0 Reference Manual: Full-Text Stopwords. Accessed from: https://dev.mysql.com/doc/refman/8.0/en/fulltext-stopwords.html on: 14/05/2018

Owens, M.J., Morgan, W.N., Plott, S.J. and Nemeroff, C.B., 1997. Neurotransmitter receptor and transporter binding profile of antidepressants and their metabolites. *Journal of Pharmacology and Experimental Therapeutics*, *283*(3), pp.1305-1322.

Overstreet, D.H., Miller, C.S., Janowsky, D.S. and Russell, R.W., 1996. Potential animal model of multiple chemical sensitivity with cholinergic supersensitivity. *Toxicology*, *111*(1), pp.119-134.

Overstreet, D. H., & Wegener, G. (2013). The flinders sensitive line rat model of depression—25 years and still producing. *Pharmacological reviews*, *65*(1), 143-155.

Oxman, A.D. and Guyatt, G.H., 1991. Validation of an index of the quality of review articles. *Journal of clinical epidemiology*, *44*(11), pp.1271-1278.

Paizanis, E., Hamon, M. and Lanfumey, L., 2007. Hippocampal neurogenesis, depressive disorders, and antidepressant therapy. *Neural plasticity*, *2007*.

Parsons, N.R., Teare, M.D. and Sitch, A.J., 2018. Unit of analysis issues in laboratory-based research. *eLife*, *7*, p.e32486.

Patten, S.B., 2006. A major depression prognosis calculator based on episode duration. *Clinical Practice and Epidemiology in Mental Health*, *2*(1), p.13.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), pp.2825-2830.

Pencina, M.J., D'Agostino Sr, R.B., D'Agostino Jr, R.B. and Vasan, R.S., 2008. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*, *27*(2), pp.157-172.

Perez-Cornago, A., Sanchez-Villegas, A., Bes-Rastrollo, M., Gea, A., Molero, P., Lahortiga-Ramos, F. and Martínez-González, M.A., 2016. Intake of High-Fat Yogurt, but Not of Low-Fat Yogurt or Prebiotics, Is Related to Lower Risk of Depression in Women of the SUN Cohort Study–3. *The Journal of nutrition*, *146*(9), pp.1731-1739.

Petrik, D., Lagace, D.C. and Eisch, A.J., 2012. The neurogenesis hypothesis of affective and anxiety disorders: are we mistaking the scaffolding for the building?. *Neuropharmacology*, *62*(1), pp.21-34.

Polter, A.M. and Kauer, J.A., 2014. Stress and VTA synapses: implications for addiction and depression. *European journal of neuroscience*, *39*(7), pp.1179-1188.

Porsolt, R.D., Anton, G., Blavet, N. and Jalfre, M., 1978. Behavioural despair in rats: a new model sensitive to antidepressant treatments. *European journal of pharmacology*, *47*(4), pp.379-391.

Pound, P. and Bracken, M.B., 2014. Is animal research sufficiently evidence based to be a cornerstone of biomedical research?. *BMJ*, *348*, pp.3387.

Pound, P., Ebrahim, S., Sandercock, P., Bracken, M.B. and Roberts, I., 2004. Where is the evidence that animal research benefits humans?. *BMJ*, *328*(7438), pp.514-517.

Prescott, M.J. and Lidster, K., 2017. Improving quality of science through better animal welfare: the NC3Rs strategy. *Lab animal*, *46*(4), p.152.

Rada, P., Avena, N.M. and Hoebel, B.G., 2005. Daily bingeing on sugar repeatedly releases dopamine in the accumbens shell. *Neuroscience*, *134*(3), pp.737-744.

Rathbone, J., Hoffmann, T. and Glasziou, P., 2015. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. *Systematic reviews*, *4*(1), p.80.

Regier, D.A., Narrow, W.E., Clarke, D.E., Kraemer, H.C., Kuramoto, S.J., Kuhl, E.A. and Kupfer, D.J., 2013. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *American journal of psychiatry*, *170*(1), pp.59-70.

Peirson, S.N., Brown, L.A., Pothecary, C.A., Benson, L.A. and Fisk, A.S., 2018. Light and the laboratory mouse. *Journal of neuroscience methods*, *300*, pp.26-36.

Ridder, S., Chourbaji, S., Hellweg, R., Urani, A., Zacher, C., Schmid, W., Zink, M., Hörtnagl, H., Flor, H., Henn, F.A. and Schütz, G., 2005. Mice with genetically altered glucocorticoid receptor expression show altered sensitivity for stress-induced depressive reactions. *Journal of Neuroscience*, *25*(26), pp.6243-6250.

Robin, X. 2017. “pROC” Package. Retrieved from: https://cran.r-project.org/web/packages/pROC/pROC.pdf

Rodriguez, J.D., Perez, A. and Lozano, J.A., 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, *32*(3), pp.569-575.

Romijn, A.R. and Rucklidge, J.J., 2015. Systematic review of evidence to support the theory of psychobiotics. *Nutrition reviews*, *73*(10), pp.675-693.

Rooke, E.D., Vesterinen, H.M., Sena, E.S., Egan, K.J. and Macleod, M.R., 2011. Dopamine agonists in animal models of Parkinson’s disease: a systematic review and meta-analysis. *Parkinsonism & related disorders*, *17*(5), pp.313-320.

Rosenthal, R. and DiMatteo, M.R., 2001. Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual review of psychology*, *52*(1), pp.59-82.

Ross, J.S., 2016. Clinical research data sharing: what an open science world means for researchers involved in evidence synthesis. *Systematic reviews*, *5*(1), p.159.

Rush, A.J., Trivedi, M.H., Wisniewski, S.R., Nierenberg, A.A., Stewart, J.W., Warden, D., Niederehe, G., Thase, M.E., Lavori, P.W., Lebowitz, B.D. and McGrath, P.J., 2006. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. *American Journal of Psychiatry*, *163*(11), pp.1905-1917.

Rush, A.J, 2007. The Varied Clinical Presentations of Major Depressive Disorder. *The Journal of clinical psychiatry*, 68(suppl 8) pp. 4-10

Rush, A.J., Trivedi, M.H., Stewart, J.W., Nierenberg, A.A., Fava, M., Kurian, B.T., Warden, D., Morris, D.W., Luther, J.F., Husain, M.M. and Cook, I.A., 2011. Combining medications to enhance depression outcomes (CO-MED): acute and long-term outcomes of a single-blind randomized study. *American Journal of Psychiatry*, *168*(7), pp.689-701.

Russell, W.M.S., Burch, R.L. and Hume, C.W., 1959. *The principles of humane experimental technique* (Vol. 238). London: Methuen.

Sampath, D., Sathyanesan, M. and Newton, S.S., 2017. Cognitive dysfunction in major depression and Alzheimer’s disease is associated with hippocampal–prefrontal cortex dysconnectivity. *Neuropsychiatric disease and treatment*, *13*, p.1509.

Satterthwaite, T.D., Kable, J.W., Vandekar, L., Katchmar, N., Bassett, D.S., Baldassano, C.F., Ruparel, K., Elliott, M.A., Sheline, Y.I., Gur, R.C. and Gur, R.E., 2015. Common and dissociable dysfunction of the reward system in bipolar and unipolar depression. *Neuropsychopharmacology*, *40*(9), p.2258-68.

Savignac, H.M., Corona, G., Mills, H., Chen, L., Spencer, J.P., Tzortzis, G. and Burnet, P.W., 2013. Prebiotic feeding elevates central brain derived neurotrophic factor, N-methyl-D-aspartate receptor subunits and D-serine. *Neurochemistry international*, *63*(8), pp.756-764.

Seretny, M., Currie, G.L., Sena, E.S., Ramnarine, S., Grant, R., MacLeod, M.R., Colvin, L.A. and Fallon, M., 2014. Incidence, prevalence, and predictors of chemotherapy-induced peripheral neuropathy: a systematic review and meta-analysis. *PAIN*, *155*(12), pp.2461-2470.

Seminowicz, D.A., Mayberg, H.S., McIntosh, A.R., Goldapple, K., Kennedy, S., Segal, Z. and Rafi-Tari, S., 2004. Limbic–frontal circuitry in major depression: a path modeling metanalysis. *Neuroimage*, *22*(1), pp.409-418.

Sena, E., van der Worp, H.B., Howells, D. and Macleod, M., 2007. How can we improve the pre-clinical development of drugs for stroke?. *Trends in neurosciences*, *30*(9), pp.433-439.

Sena, E.S., Van Der Worp, H.B., Bath, P.M., Howells, D.W. and Macleod, M.R., 2010. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS biology*, *8*(3), p.e1000344.

Sena, E.S., Currie, G.L., McCann, S.K., Macleod, M.R. and Howells, D.W., 2014. Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically. *Journal of Cerebral Blood Flow & Metabolism*, *34*(5), pp.737-742.

Shaw, D.M., Camps, F.E. and Eccleston, E.G., 1967. 5-Hydroxytryptamine in the hind-brain of depressive suicides. *The British Journal of Psychiatry*, *113*(505), pp.1407-1411.

Shea, B.J., Hamel, C., Wells, G.A., Bouter, L.M., Kristjansson, E., Grimshaw, J., Henry, D.A. and Boers, M., 2009. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of clinical epidemiology*, *62*(10), pp.1013-1020.

Shemilt, I., Simon, A., Hollands, G.J., Marteau, T.M., Ogilvie, D., O'Mara-Eves, A., Kelly, M.P. and Thomas, J., 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods*, *5*(1), pp.31-49.

Sherwin, E., Sandhu, K.V., Dinan, T.G. and Cryan, J.F., 2016. May the force be with you: the light and dark sides of the microbiota–gut–brain axis in neuropsychiatry. *CNS drugs*, *30*(11), pp.1019-1041.

Shinohara, K., Honyashiki, M., Imai, H., Hunot, V., Caldwell, D.M., Davies, P., Moore, T.H., Furukawa, T.A. and Churchill, R., 2013. Behavioural therapies versus other psychological therapies for depression. *Cochrane Database of Systematic Reviews*, (10).

Shulman, K.I., Herrmann, N. and Walker, S.E., 2013. Current place of monoamine oxidase inhibitors in the treatment of depression. *CNS drugs*, *27*(10), pp.789-797.

Sievert, C., (2018) plotly for R. Version 4.7.1. https://plotly-book.cpsievert.me

Slattery, D.A. and Cryan, J.F., 2014. The ups and downs of modelling mood disorders in rodents. *ILAR journal*, *55*(2), pp.297-309.

Souery, D., Papakostas, G.I. and Trivedi, M.H., 2006. Treatment-resistant depression. *The Journal of clinical psychiatry*, *67*, pp.16-22.

Spielmans, G.I., Berman, M.I. and Usitalo, A.N., 2011. Psychotherapy versus second-generation antidepressants in the treatment of depression: a meta-analysis. *The Journal of nervous and mental disease*, *199*(3), pp.142-149.

Steru, L., Chermat, R., Thierry, B. and Simon, P., 1985. The tail suspension test: a new method for screening antidepressants in mice. *Psychopharmacology*, *85*(3), pp.367-370.

Systematic Review Facility (SyRF). www.syrf.org.uk

Takanaga, H., Ohtsuki, S., Hosoya, K.I. and Terasaki, T., 2001. GAT2/BGT-1 as a system responsible for the transport of γ-aminobutyric acid at the mouse blood–brain barrier. Journal of Cerebral Blood Flow & Metabolism, 21(10), pp.1232-1239.

Tatsumi, M., Groshan, K., Blakely, R.D. and Richelson, E., 1997. Pharmacological profile of antidepressants and related compounds at human monoamine transporters. *European journal of pharmacology*, *340*(2-3), pp.249-258.

Thomas, J., Brunton, J., Graziosi, S., 2010. EPPI-Reviewer 4.0: software for research synthesis. EPPI-Centre Software. London: Social Science Research Unit, Institute of Education.

Tillmann, S., Awwad, H.M., Eskelund, A.R., Treccani, G., Geisel, J., Wegener, G. and Obeid, R., 2018. Probiotics Affect One-Carbon Metabolites and Catecholamines in a Genetic Rat Model of Depression. *Molecular nutrition & food research*, *62*(7), p.1701070.

Tillmann, S. and Wegener, G., 2018. Syringe-feeding as a novel delivery method for accurate individual dosing of probiotics in rats. *Beneficial microbes*, *9*(2), pp.311-315.

Thomas, L., Kessler, D., Campbell, J., Morrison, J., Peters, T.J., Williams, C., Lewis, G. and Wiles, N., 2013. Prevalence of treatment-resistant depression in primary care: cross-sectional data. *Br J Gen Pract*, *63*(617), pp.e852-e858.

Thompson, R.S., Roller, R., Mika, A., Greenwood, B.N., Knight, R., Chichlowski, M., Berg, B.M. and Fleshner, M., 2017. Dietary prebiotics and bioactive milk fractions improve NREM sleep, enhance REM sleep rebound and attenuate the stress-induced decrease in diurnal temperature and gut microbial alpha diversity. *Frontiers in behavioral neuroscience*, *10*, p.240.

Thompson, S.G. and Sharp, S.J., 1999. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in medicine*, *18*(20), pp.2693-2708.

Todd, A., Anderson, R.J. and Groundwater, P.W., 2009. Rational drug design-designing a molecule that binds to a target. *The pharmaceutical journal.*, *283*(7563), pp.131-132.

Tsafnat, G., Glasziou, P., Choong, M.K., Dunn, A., Galgani, F. and Coiera, E., 2014. Systematic review automation technologies. *Systematic reviews*, *3*(1), p.74.

Tsuruoka, Y., Tateishi, Y., Kim, J.D., Ohta, T., McNaught, J., Ananiadou, S. and Tsujii, J.I., 2005, November. Developing a robust part-of-speech tagger for biomedical text. In *Panhellenic Conference on Informatics* (pp. 382-392). Springer, Berlin, Heidelberg.

Tuunainen, A., Kripke, D.F. and Endo, T., 2004. Light therapy for non-seasonal depression. *Cochrane Database Syst Rev*, *2*(2).

Tye, K.M., Mirzabekov, J.J., Warden, M.R., Ferenczi, E.A., Tsai, H.C., Finkelstein, J., Kim, S.Y., Adhikari, A., Thompson, K.R., Andalman, A.S. and Gunaydin, L.A., 2013. Dopamine neurons modulate neural encoding and expression of depression-related behaviour. *Nature*, *493*(7433), p.537.

UNESCO, 2017. Global Open Access Portal: Open Science Movement. Retrieved from: http://www.unesco.org/new/en/communication-and-information/portals-and-platforms/goap/open-science-movement/

Van der Worp, H.B., Howells, D.W., Sena, E.S., Porritt, M.J., Rewell, S., O'Collins, V. and Macleod, M.R., 2010. Can animal models of disease reliably inform human studies?. *PLoS medicine*, *7*(3), p.e1000245.

van der Worp, H.B., Sena, E.S., Donnan, G.A., Howells, D.W. and Macleod, M.R., 2007. Hypothermia in animal models of acute ischaemic stroke: a systematic review and meta-analysis. *Brain*, *130*(12), pp.3063-3074.

Van Riezen, H., Schnieden, H. and Wren, A., 1976. Behavioural changes following olfactory bulbectomy in rats: a possible model for the detection of antidepressant drugs [proceedings]. *British journal of pharmacology*, *57*(3), p.426P.

Van Loo, H.M., De Jonge, P., Romeijn, J.W., Kessler, R.C. and Schoevers, R.A., 2012. Data-driven subtypes of major depressive disorder: a systematic review. *BMC medicine*, *10*(1), p.156.

Vesterinen, H.M., Sena, E.S., Ffrench-Constant, C., Williams, A., Chandran, S. and Macleod, M.R., 2010. Improving the translational hit of experimental treatments in multiple sclerosis. *Multiple Sclerosis Journal*, *16*(9), pp.1044-1055.

Vesterinen, H.M., Currie, G.L., Carter, S., Mee, S., Watzlawick, R., Egan, K.J., Macleod, M.R. and Sena, E.S., 2013. Systematic review and stratified meta-analysis of the efficacy of RhoA and Rho kinase inhibitors in animal models of ischaemic stroke. *Systematic reviews*, *2*(1), p.33.

Vesterinen, H.M., Sena, E.S., Egan, K.J., Hirst, T.C., Churolov, L., Currie, G.L., Antonic, A., Howells, D.W. and Macleod, M.R., 2014. Meta-analysis of data from animal studies: a practical guide. *Journal of neuroscience methods*, *221*, pp.92-102.

Vitaterna, M.H., King, D.P., Chang, A.M., Kornhauser, J.M., Lowrey, P.L., McDonald, J.D., Dove, W.F., Pinto, L.H., Turek, F.W. and Takahashi, J.S., 1994. Mutagenesis and mapping of a mouse gene, Clock, essential for circadian behavior. *Science*, *264*(5159), pp.719-725.

Wagner, A.B., 2010. Open access citation advantage: An annotated bibliography. *Issues in Science and Technology Librarianship, Winter,* DOI: 10.5062/F4Q81B0W

Walker, F.R., James, M.H., Hickie, I.B. and McGorry, P.D., 2014. Clinical staging: a necessary step in the development of improved animal models of mood disturbance?. *International Journal of Neuropsychopharmacology*, *17*(3), pp.491-495.

Wallace, B.C., Small, K., Brodley, C.E. and Trikalinos, T.A., 2010, July. Active learning for biomedical citation screening. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 173-182). ACM. (A)

Wallace, B.C., Trikalinos, T.A., Lau, J., Brodley, C. and Schmid, C.H., 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, *11*(1), p.55. (B)

Wallace, B.C., Small, K., Brodley, C.E., Lau, J., Schmid, C.H., Bertram, L., Lill, C.M., Cohen, J.T. and Trikalinos, T.A., 2012. Toward modernizing the systematic review

pipeline in genetics: efficient updating via data mining. *Genetics in medicine*, *14*(7), p.663. (A)

Wallace, B.C., Small, K., Brodley, C.E., Lau, J. and Trikalinos, T.A., 2012, January. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium* (pp. 819-824). ACM. (B)

Wang, Q., Timberlake II, M.A., Prall, K. and Dwivedi, Y., 2017. The recent progress in animal models of depression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, *77*, pp.99-109.

Wang, H., Lee, I.S., Braun, C. and Enck, P., 2016. Effect of probiotics on central nervous system functions in animals and humans: a systematic review. *Journal of neurogastroenterology and motility*, *22*(4), p.589.

Wang, Q., Liao, J., Hair, K., Bannach-Brown, A., Bahor, Z., Currie, G.L., McCann, S.K., Howells, D.W., Sena, E.S. and Macleod, M.R., 2018. Estimating the statistical performance of different approaches to meta-analysis of data from animal studies in identifying the impact of aspects of study design. *bioRxiv*, p.256776.

Watzlawick, R., Rind, J., Sena, E.S., Brommer, B., Zhang, T., Kopp, M.A., Dirnagl, U., Macleod, M.R., Howells, D.W. and Schwab, J.M., 2016. Olfactory ensheathing cell transplantation in experimental spinal cord injury: effect size and reporting bias of 62 experimental treatments: a systematic review and meta-analysis. *PLoS biology*, *14*(5), p.e1002468.

Webb, G.I., Lee, L.K., Petitjean, F. and Goethals, B., 2017. Understanding Concept Drift. *arXiv preprint arXiv:1704.00362*.

Whiteford, H.A., Degenhardt, L., Rehm, J., Baxter, A.J., Ferrari, A.J., Erskine, H.E., Charlson, F.J., Norman, R.E., Flaxman, A.D., Johns, N. and Burstein, R., 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. The Lancet, 382(9904), pp.1575-1586.

Wilkins, D., (2018) 'treemapify' R Package. https://CRAN.R-project.org/package=treemapify

Williams, N.C., Johnson, M.A., Shaw, D.E., Spendlove, I., Vulevic, J., Sharpe, G.R. and Hunter, K.A., 2016. A prebiotic galactooligosaccharide mixture reduces severity of hyperpnoea-induced bronchoconstriction and markers of airway inflammation. *British Journal of Nutrition*, *116*(5), pp.798-804.

Williams, N.R. and Schatzberg, A.F., 2016. NMDA antagonist treatment of depression. *Current opinion in neurobiology*, *36*, pp.112-117.

Willner, P., 1984. The validity of animal models of depression. *Psychopharmacology*, *83*(1), pp.1-16.

Willner, P. and Mitchell, P.J., 2002. The validity of animal models of predisposition to depression. *Behavioural pharmacology*, *13*(3), pp.169-188.

World Health Organization, 2017. Depression and other common mental disorders: global health estimates. Accessed from: http://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf;jsessionid=465EEAACC60B7941CAE57A695979197C?sequence=1

World Health Organization. (2018). *International statistical classification of diseases and related health problems* (11th Revision). Retrieved from https://icd.who.int/browse11/l-m/en

Wray, N.R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E.M., Abdellaoui, A., Adams, M.J., Agerbo, E., Air, T.M., Andlauer, T.M. and Bacanu, S.A., 2018. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics*, *50*(5), p.668.

Yang, Y., Cui, Y., Sang, K., Dong, Y., Ni, Z., Ma, S. and Hu, H., 2018. Ketamine blocks bursting in the lateral habenula to rapidly relieve depression. *Nature*, *554*(7692), p.317.

Yassa, M.A. and Stark, C.E., 2011. Pattern separation in the hippocampus. *Trends in neurosciences*, *34*(10), pp.515-525.

Youdim, M.B., Edmondson, D. and Tipton, K.F., 2006. The therapeutic potential of monoamine oxidase inhibitors. *Nature Reviews Neuroscience*, *7*(4), p.295.

Zhang, Y., Marshall, I. and Wallace, B.C., 2016, November. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (Vol. 2016, p. 795). NIH Public Access.

Zhao, Y., Ma, R., Shen, J., Su, H., Xing, D. and Du, L., 2008. A mouse model of depression induced by repeated corticosterone injections. *European journal of pharmacology*, *581*(1-2), pp.113-120.

Zhu, X., Wang, X., Xiao, J., Liao, J., Zhong, M., Wang, W. and Yao, S., 2012. Evidence of a dissociation pattern in resting-state default mode network connectivity in first-episode, treatment-naive major depression patients. *Biological psychiatry*, *71*(7), pp.611-617.

Zohar, J., Nutt, D.J., Kupfer, D.J., Moller, H.J., Yamawaki, S., Spedding, M. and Stahl, S.M., 2014. A proposal for an updated neuropsychopharmacological nomenclature. *European Neuropsychopharmacology*, *24*(7), pp.1005-1014.

Zucker, I. and Beery, A.K., 2010. Males still dominate animal studies. *Nature*, *465*(7299), p.690.

Zwetsloot, P.P., Van Der Naald, M., Sena, E.S., Howells, D.W., IntHout, J., De Groot, J.A., Chamuleau, S.A., MacLeod, M.R. and Wever, K.E., 2017. Standardized mean differences cause funnel plot distortion in publication bias assessments. *elife*, *6*, p.e24260.

REFERENCES IN SYSTEMATIC REVIEW CHAPTER 5: INTERVENTIONS TARGETING THE GUT MICROBIOTA IN ANIMAL MODELS OF DEPRESSION: A SYSTEMATIC REVIEW AND META-ANALYSIS

Ait-Belgnaoui, A., Colom, A., Braniste, V., Ramalho, L., Marrot, A., Cartier, C., Houdeau, E., Theodorou, V. and Tompkins, T., 2014. Probiotic gut effect prevents the chronic psychological stress-induced brain activity abnormality in mice. *Neurogastroenterology & Motility*, *26*(4), pp.510-520.

Bravo, J.A., Forsythe, P., Chew, M.V., Escaravage, E., Savignac, H.M., Dinan, T.G., Bienenstock, J. and Cryan, J.F., 2011. Ingestion of Lactobacillus strain regulates emotional behavior and central GABA receptor expression in a mouse via the vagus nerve. *Proceedings of the National Academy of Sciences*, p.201102999.

Desbonnet, L., Garrett, L., Clarke, G., Kiely, B., Cryan, J.F. and Dinan, T.G., 2010. Effects of the probiotic Bifidobacterium infantis in the maternal separation model of depression. *Neuroscience*, *170*(4), pp.1179-1188.

Farshim, P., Walton, G., Chakrabarti, B., Givens, I., Saddy, D., Kitchen, I., Swann, J.R. and Bailey, A., 2016. Maternal weaning modulates emotional behavior and regulates the gut-brain axis. *Scientific reports*, *6*, p.21958.

Gacias, M., Gaspari, S., Santos, P.M.G., Tamburini, S., Andrade, M., Zhang, F., Shen, N., Tolstikov, V., Kiebish, M.A., Dupree, J.L. and Zachariou, V., 2016. Microbiota-driven transcriptional changes in prefrontal cortex override genetic differences in social behavior. *elife*, *5*, p.e13442.

Gárate, I., García-Bueno, B., Madrigal, J.L., Bravo, L., Berrocoso, E., Caso, J.R., Micó, J.A. and Leza, J.C., 2011. Origin and consequences of brain Toll-like receptor 4 pathway stimulation in an experimental model of depression. *Journal of neuroinflammation*, *8*(1), p.151.

Ilgin, S., Can, O.D., Atli, O., Ucel, U.I., Sener, E. and Guven, I., 2015. Ciprofloxacin-induced neurotoxicity: evaluation of possible underlying mechanisms. *Toxicology mechanisms and methods*, *25*(5), pp.374-381.

Jørgensen, B.P., Hansen, J.T., Krych, L., Larsen, C., Klein, A.B., Nielsen, D.S., Josefsen, K., Hansen, A.K. and Sørensen, D.B., 2014. A possible link between food and mood: dietary impact on gut microbiota and behavior in BALB/c mice. *PloS one*, *9*(8), p.e103398.

Liang, S., Wang, T., Hu, X., Luo, J., Li, W., Wu, X., Duan, Y. and Jin, F., 2015. Administration of Lactobacillus helveticus NS8 improves behavioral, cognitive, and biochemical aberrations caused by chronic restraint stress. *Neuroscience*, *310*, pp.561-577.

Liu, Y.W., Liu, W.H., Wu, C.C., Juan, Y.C., Wu, Y.C., Tsai, H.P., Wang, S. and Tsai, Y.C., 2016. Psychotropic effects of Lactobacillus plantarum PS128 in early life-stressed and naïve adult mice. *Brain research*, *1631*, pp.1-12.

O'Mahony, S.M., Marchesi, J.R., Scully, P., Codling, C., Ceolho, A.M., Quigley, E.M., Cryan, J.F. and Dinan, T.G., 2009. Early life stress alters behavior, immunity, and microbiota in rats: implications for irritable bowel syndrome and psychiatric illnesses. *Biological psychiatry*, *65*(3), pp.263-267.

Savignac, H.M., Kiely, B., Dinan, T.G. and Cryan, J.F., 2014. Bifidobacteria exert strain-specific effects on stress-related behavior and physiology in BALB/c mice. *Neurogastroenterology & Motility*, *26*(11), pp.1615-1627.

Winther, G., Jørgensen, B.M.P., Elfving, B., Nielsen, D.S., Kihl, P., Lund, S., Sørensen, D.B. and Wegener, G., 2015. Dietary magnesium deficiency alters gut microbiota and leads to depressive-like behaviour. *Acta neuropsychiatrica*, *27*(3), pp.168-176.

Wong, M.L., Inserra, A., Lewis, M.D., Mastronardi, C.A., Leong, L., Choo, J., Kentish, S., Xie, P., Morrison, M., Wesselingh, S.L. and Rogers, G.B., 2016. Inflammasome signaling affects anxiety-and depressive-like behavior and gut microbiome composition. *Molecular psychiatry*, *21*(6), p.797.

Zheng P, Zeng B, Zhou C, Liu M, Fang Z, Xu X, Zeng L, Chen J, Fan S, Du X, Zhang X. Gut microbiome remodeling induces depressive-like behaviors through a pathway mediated by the host's metabolism. Molecular psychiatry. 2016 Jun;21(6):786.

REFERENCES IN SYSTEMATIC REVIEW CHAPTER 6: THE ANTIDEPRESSANT EFFECT OF KETAMINE IN ANIMAL MODELS OF DEPRESSION AS MEASURED BY THE FORCED SWIM TEST: A SYSTEMATIC REVIEW AND META-ANALYSIS

Akinfiresoye, L. and Tizabi, Y., 2013. Antidepressant effects of AMPA and ketamine combination: role of hippocampal BDNF, synapsin, and mTOR. *Psychopharmacology*, *230*(2), pp.291-298.

Antony, L.J., Paruchuri, V.N.K. and Ramanan, R., 2014. Antidepressant effect of ketamine in sub anaesthetic doses in male albino mice. *Journal of clinical and diagnostic research: JCDR*, *8*(6), p.HC05.

Assis, L.C., Rezin, G.T., Comim, C.M., Valvassori, S.S., Jeremias, I.C., Zugno, A.I., Quevedo, J. and Streck, E.L., 2009. Effect of acute administration of ketamine and imipramine on creatine kinase activity in the brain of rats. *Revista Brasileira de Psiquiatria*, *31*(3), pp.247-252.

Autry, A.E., Adachi, M., Nosyreva, E., Na, E.S., Los, M.F., Cheng, P.F., Kavalali, E.T. and Monteggia, L.M., 2011. NMDA receptor blockade at rest triggers rapid behavioural antidepressant responses. *Nature*, *475*(7354), p.91.

Carrier, N. and Kabbaj, M., 2013. Sex differences in the antidepressant-like effects of ketamine. *Neuropharmacology*, *70*, pp.27-34.

Chiu, C.T., Scheuing, L., Liu, G., Liao, H.M., Linares, G.R., Lin, D. and Chuang, D.M., 2015. The mood stabilizer lithium potentiates the antidepressant-like effects and ameliorates oxidative stress induced by acute ketamine in a mouse model of stress. *International Journal of Neuropsychopharmacology*, *18*(6), p.pyu102.

Cruz, S.L., Soberanes-Chávez, P., Páez-Martinez, N. and López-Rubalcava, C., 2009. Toluene has antidepressant-like actions in two animal models used for the screening of antidepressant drugs. *Psychopharmacology*, *204*(2), pp.279-286.

da Silva, F.C.C., de Oliveira Cito, M.D.C., da Silva, M.I.G., Moura, B.A., de Aquino Neto, M.R., Feitosa, M.L., de Castro Chaves, R., Macedo, D.S., de Vasconcelos, S.M.M., de França Fonteles, M.M. and de Sousa, F.C.F., 2010. Behavioral alterations and pro-oxidant effect of a single ketamine administration to mice. *Brain research bulletin*, *83*(1-2), pp.9-15.

da Silva Moreira, S.F., Nunes, E.A., Kuo, J., de Macedo, I.C., Muchale, A., de Oliveira, C., Scarabelot, V.L., Marques Filho, P.R., Medeiros, L.F., Caumo, W. and Torres, I.L., 2016. Hypoestrogenism alters mood: ketamine reverses depressive-like behavior induced by ovariectomy in rats. *Pharmacological Reports*, *68*(1), pp.109-115.

Engin, E., Treit, D. and Dickson, C.T., 2009. Anxiolytic-and antidepressant-like properties of ketamine in behavioral and neurophysiological animal models. *Neuroscience*, *161*(2), pp.359-369.

Franceschelli, A., Sens, J., Herchick, S., Thelen, C. and Pitychoutis, P.M., 2015. Sex differences in the rapid and the sustained antidepressant-like effects of

ketamine in stress-naive and “depressed” mice exposed to chronic mild stress. *Neuroscience*, *290*, pp.49-60.

Fuchikami, M., Thomas, A., Liu, R., Wohleb, E.S., Land, B.B., DiLeone, R.J., Aghajanian, G.K. and Duman, R.S., 2015. Optogenetic stimulation of infralimbic PFC reproduces ketamine’s rapid and sustained antidepressant actions. *Proceedings of the National Academy of Sciences*, *112*(26), pp.8106-8111.

Garcia, L.S., Comim, C.M., Valvassori, S.S., Réus, G.Z., Barbosa, L.M., Andreazza, A.C., Stertz, L., Fries, G.R., Gavioli, E.C., Kapczinski, F. and Quevedo, J., 2008. Acute administration of ketamine induces antidepressant-like effects in the forced swimming test and increases BDNF levels in the rat hippocampus. *Progress in neuro-psychopharmacology and biological psychiatry*, *32*(1), pp.140-144.

Garcia, L.S., Comim, C.M., Valvassori, S.S., Réus, G.Z., Andreazza, A.C., Stertz, L., Fries, G.R., Gavioli, E.C., Kapczinski, F. and Quevedo, J., 2008. Chronic administration of ketamine elicits antidepressant-like effects in rats without affecting hippocampal brain-derived neurotrophic factor protein levels. *Basic & clinical pharmacology & toxicology*, *103*(6), pp.502-506.

Ghasemi, M., Kazemi, M.H., Yoosefi, A., Ghasemi, A., Paragomi, P., Amini, H. and Afzali, M.H., 2014. Rapid antidepressant effects of repeated doses of ketamine compared with electroconvulsive therapy in hospitalized patients with major depressive disorder. *Psychiatry Research*, *215*(2), pp.355-361.

Haj-Mirzaian, A., Amiri, S., Kordjazy, N., Rahimi-Balaei, M., Haj-Mirzaian, A., Marzban, H., Aminzadeh, A., Dehpour, A.R. and Mehr, S.E., 2015. Blockade of NMDA receptors reverses the depressant, but not anxiogenic effect of adolescence social isolation in mice. *European journal of pharmacology*, *750*, pp.160-166.

Jett, J.D., Boley, A.M., Girotti, M., Shah, A., Lodge, D.J. and Morilak, D.A., 2015. Antidepressant-like cognitive and behavioral effects of acute ketamine administration associated with plasticity in the ventral hippocampus to medial prefrontal cortex pathway. *Psychopharmacology*, *232*(17), pp.3123-3133.

Kilic, F.S., Ismailoglu, S., Kaygisiz, B. and Oner, S., 2014. Effects of single and combined gabapentin use in elevated plus maze and forced swimming tests. *Acta neuropsychiatrica*, *26*(5), pp.307-314.

Koike, H. and Chaki, S., 2014. Requirement of AMPA receptor stimulation for the sustained antidepressant activity of ketamine and LY341495 during the forced swim test in rats. *Behavioural brain research*, *271*, pp.111-115.

Koike, H., Iijima, M. and Chaki, S., 2013. Effects of ketamine and LY341495 on the depressive-like behavior of repeated corticosterone-injected rats. *Pharmacology Biochemistry and Behavior*, *107*, pp.20-23.

Li, N., Liu, R.J., Dwyer, J.M., Banasr, M., Lee, B., Son, H., Li, X.Y., Aghajanian, G. and Duman, R.S., 2011. Glutamate N-methyl-D-aspartate receptor antagonists rapidly reverse behavioral and synaptic deficits caused by chronic stress exposure. *Biological psychiatry*, *69*(8), pp.754-761.

Li, Y., Zhu, Z.R., Ou, B.C., Wang, Y.Q., Tan, Z.B., Deng, C.M., Gao, Y.Y., Tang, M., So, J.H., Mu, Y.L. and Zhang, L.Q., 2015. Dopamine D2/D3 but not dopamine D1 receptors are involved in the rapid antidepressant-like effects of ketamine in the forced swim test. *Behavioural brain research*, *279*, pp.100-105.

Liebenberg, N., Joca, S. and Wegener, G., 2015. Nitric oxide involvement in the antidepressant-like effect of ketamine in the Flinders sensitive line rat model of depression. *Acta neuropsychiatrica*, *27*(2), pp.90-96.

Ma, X.C., Dang, Y.H., Jia, M., Ma, R., Wang, F., Wu, J., Gao, C.G. and Hashimoto, K., 2013. Long-lasting antidepressant action of ketamine, but not glycogen synthase kinase-3 inhibitor SB216763, in the chronic mild stress model of mice. *PLoS One*, *8*(2), p.e56053.

Młyniec, K., Budziszewska, B., Holst, B., Ostachowicz, B. and Nowak, G., 2015. GPR39 (zinc receptor) knockout mice exhibit depression-like behavior and CREB/BDNF down-regulation in the hippocampus. *International Journal of Neuropsychopharmacology*, *18*(3).

Moskal, J.R., Burch, R., Burgdorf, J.S., Kroes, R.A., Stanton, P.K., Disterhoft, J.F. and Leander, J.D., 2014. GLYX-13, an NMDA receptor glycine site functional partial agonist enhances cognition and produces antidepressant effects without the psychotomimetic side effects of NMDA receptor antagonists. *Expert opinion on investigational drugs*, *23*(2), pp.243-254.

Nosyreva, E., Autry, A.E., Kavalali, E.T. and Monteggia, L.M., 2014. Age dependence of the rapid antidepressant and synaptic effects of acute NMDA receptor blockade. *Frontiers in molecular neuroscience*, *7*, p.94.

Perrine, Shane A., Farhad Ghoddoussi, Mark S. Michaels, Imran S. Sheikh, George McKelvey, and Matthew P. Galloway. "Ketamine reverses stress-induced depression-like behavior and increased GABA levels in the anterior cingulate: an 11.7 T 1H-MRS study in rats." *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 51 (2014): 9-15.

Parise, E.M., Alcantara, L.F., Warren, B.L., Wright, K.N., Hadad, R., Sial, O.K., Kroeck, K.G., Iñiguez, S.D. and Bolaños-Guzmán, C.A., 2013. Repeated ketamine exposure induces an enduring resilient phenotype in adolescent and adult rats. *Biological psychiatry*, *74*(10), pp.750-759.

Petryshen, T.L., Lewis, M.C., Dennehy, K.A., Garza, J.C. and Fava, M., 2016. Antidepressant-like effect of low dose ketamine and scopolamine co-treatment in mice. *Neuroscience letters*, *620*, pp.70-73.

Popik, P., Kos, T., Sowa-Kućma, M. and Nowak, G., 2008. Lack of persistent effects of ketamine in rodent models of depression. *Psychopharmacology*, *198*(3), pp.421-430.

Pozzi, L., Dorocic, I.P., Wang, X., Carlén, M. and Meletis, K., 2014. Mice lacking NMDA receptors in parvalbumin neurons display normal depression-related behavior and response to antidepressant action of NMDAR antagonists. *PloS one*, *9*(1), p.e83879.

Prabhakar, A., H.S. Somashekar, C.G. Gokul, R. Santosh, AbhishekAcharya and K.M. Naveen, 2011. Antidepressant activity of ketamine in albino mice. *Pharmacologyonline, 2:* 241-252.

Ren, Z., Pribiag, H., Jefferson, S.J., Shorey, M., Fuchs, T., Stellwagen, D. and Luscher, B., 2016. Bidirectional homeostatic regulation of a depression-related brain state by gamma-aminobutyric acidergic deficits and ketamine treatment. *Biological psychiatry*, *80*(6), pp.457-468.

Ren, Q., Ma, M., Yang, C., Zhang, J.C., Yao, W. and Hashimoto, K., 2015. BDNF–TrkB signaling in the nucleus accumbens shell of mice has key role in methamphetamine withdrawal symptoms. *Translational psychiatry*, *5*(10), p.e666.

Réus, G.Z., Stringari, R.B., Ribeiro, K.F., Ferraro, A.K., Vitto, M.F., Cesconetto, P., Souza, C.T. and Quevedo, J., 2011. Ketamine plus imipramine treatment induces antidepressant-like behavior and increases CREB and BDNF protein levels and PKA and PKC phosphorylation in rat brain. *Behavioural brain research*, *221*(1), pp.166-171.

Réus, G.Z., Abelaira, H.M., dos Santos, M.A.B., Carlessi, A.S., Tomaz, D.B., Neotti, M.V., Liranço, J.L.G., Gubert, C., Barth, M., Kapczinski, F. and Quevedo, J., 2013. Ketamine and imipramine in the nucleus accumbens regulate histone deacetylation induced by maternal deprivation and are critical for associated behaviors. *Behavioural brain research*, *256*, pp.451-456.

Réus, G.Z., Vieira, F.G., Abelaira, H.M., Michels, M., Tomaz, D.B., dos Santos, M.A.B., Carlessi, A.S., Neotti, M.V., Matias, B.I., Luz, J.R. and Dal-Pizzol, F., 2014. MAPK signaling correlates with the antidepressant effects of ketamine. *Journal of psychiatric research*, *55*, pp.15-21.

Réus, G.Z., Carlessi, A.S., Titus, S.E., Abelaira, H.M., Ignácio, Z.M., da Luz, J.R., Matias, B.I., Bruchchen, L., Florentino, D., Vieira, A. and Petronilho, F., 2015. A single dose of S-ketamine induces long-term antidepressant effects and decreases oxidative stress in adulthood rats following maternal deprivation. *Developmental neurobiology*, *75*(11), pp.1268-1281.

Réus, G.Z., Nacif, M.P., Abelaira, H.M., Tomaz, D.B., dos Santos, M.A.B., Carlessi, A.S., da Luz, J.R., Gonçalves, R.C., Vuolo, F., Dal-Pizzol, F. and Carvalho, A.F., 2015. Ketamine ameliorates depressive-like behaviors and immune alterations in adult rats following maternal deprivation. *Neuroscience letters*, *584*, pp.83-87.

Sałat, K., Siwek, A., Starowicz, G., Librowski, T., Nowak, G., Drabik, U., Gajdosz, R. and Popik, P., 2015. Antidepressant-like effects of ketamine, norketamine and dehydronorketamine in forced swim test: role of activity at NMDA receptor. *Neuropharmacology*, *99*, pp.301-307.

Sarkar, A. and Kabbaj, M., 2016. Sex differences in effects of ketamine on behavior, spine density, and synaptic proteins in socially isolated rats. *Biological psychiatry*, *80*(6), pp.448-456.

Saur, L., Bagatini, P.B., Greggio, S., Venturin, G.T., Vaz, S.P., dos Reis Ferreira, K., Junqueira, J.S., Lara, D.R., DaCosta, J.C., Jeckel, C.M.M. and Mestriner, R.G., 2015. Antidepressant Effects of Ketamine Are Not Related to 18 F-FDG Metabolism

or Tyrosine Hydroxylase Immunoreactivity in the Ventral Tegmental Area of Wistar Rats. *Neurochemical research*, *40*(6), pp.1153-1164.

Sun, H.L., Zhou, Z.Q., Zhang, G.F., Yang, C., Wang, X.M., Shen, J.C., Hashimoto, K. and Yang, J.J., 2017. Role of hippocampal p11 in the sustained antidepressant effect of ketamine in the chronic unpredictable mild stress model. *Translational psychiatry*, *6*(2), p.e741.

Tang, J., Xue, W., Xia, B., Ren, L., Tao, W., Chen, C., Zhang, H., Wu, R., Wang, Q., Wu, H. and Duan, J., 2015. Involvement of normalized NMDA receptor and mTOR-related signaling in rapid antidepressant effects of Yueju and ketamine on chronically stressed mice. *Scientific reports*, *5*, p.13573.

Tizabi, Y., Bhatti, B.H., Manaye, K.F., Das, J.R. and Akinfiresoye, L., 2012. Antidepressant-like effects of low ketamine dose is associated with increased hippocampal AMPA/NMDA receptor density ratio in female Wistar–Kyoto rats. *Neuroscience*, *213*, pp.72-80.

Vogt, M.A., Vogel, A.S., Pfeiffer, N., Gass, P. and Inta, D., 2015. Role of the nitric oxide donor sodium nitroprusside in the antidepressant effect of ketamine in mice. *European Neuropsychopharmacology*, *25*(10), pp.1848-1852.

Walker, A.K., Budac, D.P., Bisulco, S., Lee, A.W., Smith, R.A., Beenders, B., Kelley, K.W. and Dantzer, R., 2013. NMDA receptor blockade by ketamine abrogates lipopolysaccharide-induced depressive-like behavior in C57BL/6J mice. *Neuropsychopharmacology*, *38*(9), p.1609.

Walker, A.J., Foley, B.M., Sutor, S.L., McGillivray, J.A., Frye, M.A. and Tye, S.J., 2015. Peripheral proinflammatory markers associated with ketamine response in a preclinical model of antidepressant-resistance. *Behavioural brain research*, *293*, pp.198-202.

Wang, J., Goffer, Y., Xu, D., Tukey, D.S., Shamir, D.B., Eberle, S.E., Zou, A.H., Blanck, T.J. and Ziff, E.B., 2011. A single subanesthetic dose of ketamine relieves depression-like behaviors induced by neuropathic pain in rats. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, *115*(4), pp.812-821.

Wang, N., Yu, H.Y., Shen, X.F., Gao, Z.Q., Yang, C., Yang, J.J. and Zhang, G.F., 2015. The rapid antidepressant effect of ketamine in rats is associated with down-regulation of pro-inflammatory cytokines in the hippocampus. *Upsala journal of medical sciences*, *120*(4), pp.241-248.

Wróbel, A., Serefko, A., Wlaź, P. and Poleszak, E., 2015. The effect of imipramine, ketamine, and zinc in the mouse model of depression. *Metabolic brain disease*, *30*(6), pp.1379-1386.

Xia, B., Zhang, H., Xue, W., Tao, W., Chen, C., Wu, R., Ren, L., Tang, J., Wu, H., Cai, B. and Doronc, R., 2016. Instant and lasting down-regulation of NR1 expression in the hippocampus is associated temporally with antidepressant activity after acute Yueju. *Cellular and molecular neurobiology*, *36*(7), pp.1189-1196.

Xu, S.X., Zhou, Z.Q., Li, X.M., Ji, M.H., Zhang, G.F. and Yang, J.J., 2013. The activation of adenosine monophosphate-activated protein kinase in rat hippocampus

contributes to the rapid antidepressant effect of ketamine. *Behavioural brain research*, *253*, pp.305-309.

Yang, C., Hu, Y.M., Zhou, Z.Q., Zhang, G.F. and Yang, J.J., 2013. Acute administration of ketamine in rats increases hippocampal BDNF and mTOR levels during forced swimming test. *Upsala journal of medical sciences*, *118*(1), pp.3-8.

Yang C, Shen J, Hong T, Hu TT, Li ZJ, Zhang HT, Zhang YJ, Zhou ZQ, Yang JJ. 2013 Effects of ketamine on lipopolysaccharide-induced depressive-like behavior and the expression of inflammatory cytokines in the rat prefrontal cortex. Molecular medicine reports. Sep 1;8(3):887-90.

Yang, C., Shirayama, Y., Zhang, J.C., Ren, Q., Yao, W., Ma, M., Dong, C. and Hashimoto, K., 2015. R-ketamine: a rapid-onset and sustained antidepressant without psychotomimetic side effects. *Translational psychiatry*, *5*(9), p.e632.

Yilmaz, A., Schulz, D., Aksoy, A. and Canbeyli, R., 2002. Prolonged effect of an anesthetic dose of ketamine on behavioral despair. *Pharmacology Biochemistry and Behavior*, *71*(1-2), pp.341-344.

Zanos, P., Piantadosi, S.C., Wu, H.Q., Pribut, H.J., Dell, M.J., Can, A., Snodgrass, H.R., Zarate, C.A., Schwarcz, R. and Gould, T.D., 2015. The prodrug 4-chlorokynurenine causes ketamine-like antidepressant effects, but not side effects, by NMDA/glycineB-site inhibition. *Journal of Pharmacology and Experimental Therapeutics*, *355*(1), pp.76-85.

Zhang, G.F., Wang, N., Shi, J.Y., Xu, S.X., Li, X.M., Ji, M.H., Zuo, Z.Y., Zhou, Z.Q. and Yang, J.J., 2013. Inhibition of the l-arginine–nitric oxide pathway mediates the antidepressant effects of ketamine in rats in the forced swimming test. *Pharmacology Biochemistry and Behavior*, *110*, pp.8-12.

Zhang, J.C., Li, S.X. and Hashimoto, K., 2014. R (−)-ketamine shows greater potency and longer lasting antidepressant effects than S (+)-ketamine. *Pharmacology Biochemistry and Behavior*, *116*, pp.137-141.

Zhang, G.F., Liu, W.X., Qiu, L.L., Guo, J., Wang, X.M., Sun, H.L., Yang, J.J. and Zhou, Z.Q., 2015. Repeated ketamine administration redeems the time lag for citalopram's antidepressant-like effects. *European Psychiatry*, *30*(4), pp.504-510.

Zhang, J.C., Yao, W., Dong, C., Yang, C., Ren, Q., Ma, M., Han, M. and Hashimoto, K., 2015. Comparison of ketamine, 7, 8-dihydroxyflavone, and ANA-12 antidepressant effects in the social defeat stress model of depression. *Psychopharmacology*, *232*(23), pp.4325-4335.

APPENDIX 1: Europe PubMed Central Search in ContentMine tool ‘getpapers’

getpapers -q '((("depressive disorder" OR "depression" OR "depressive behavior" OR "depressive behaviour" OR "dysthymia" OR "dysthymic") AND animal) NOT ("PTSD" OR "posttraumatic stress disorde"r OR "postpartum" OR "schizophrenia" OR PUB_TYPE:"comment" OR PUB_TYPE:"letter" OR PUB_TYPE:"Review"))'

APPENDIX 2: Shiny Application Code in R

Live Application available at:

https://camarades.shinyapps.io/Preclinical-Models-of-Depression/

Code available at:

https://github.com/abannachbrown/Regex-dictionary-app

APPENDIX 3: Summary Tables from Chapter 5: INTERVENTIONS TARGETING THE GUT MICROBIOTA IN ANIMAL MODELS OF DEPRESSION: A SYSTEMATIC REVIEW AND META-ANALYSIS

Table 1. Study Design Characteristics of Gut Microbiota-targeting Interventions to Reduce Depression

Authors	Year	Method of Model Induction	Intervention	Intervention Category	Dose (Units)	Route of Administration	Number of Times Administered	Treatment given Pre or Post Model Induction	Outcome Measured
Ait-Belgnaoi et al.,	2014	Water Avoidance Stress	Probio'Stick	Probiotic	1E+09 (CFU)	Oral	14 days Once daily	Pre	FOS-Immunoreactivity, Corticosterone, Adrenaline, Noradrenaline, Occulin & JAMA-A Protein Expression, Doublecortin-expressing cells, % 51-Cre-EDTA recovery
		Water Avoidance Stress	L. salivarius HA113	Probiotic	1E+09 (CFU)	Oral			
Bravo et al.,	2011	Forced Swim Test	L. rhamnosus (JB-1)	Probiotic	1E+09 (CFU)	Oral Gavage	28 days Once daily	Pre	Forced Swim Test, Corticosterone
Desbonnet et al.,	2010	Maternal Separation	Bifidobacterium infantis 35624	Probiotic	1E+10 (CFU)	Oral	45 days Once daily	Post	Forced Swim Test, Corticosterone, Noradrenaline, Dopamine, Serotonin, DOPAC, 5HIAA,

									HVA, Tryptophan, L-Kynurenine, Kynurenic acid, IL6 (LPS), IFN-Y (LPS), TNF-alpha (LPS), IL-10 (LPS)
Gacias et al.,	2016	Gastric Gavage	Antibiotic Cocktail	Antibiotic	Vancomycin (50 mg/kg), neomycin (100 mg/kg), metronidazole (100 mg/kg), amphotericin B (1 mg/kg)	Oral Gavage	21 days Once daily	Simultaneous	Forced Swim Test, Social Interaction
Gárate et al.,	2011	Chronic Mild Stress	Streptomycin Sulphate & Penicillin G	Antibiotic	Streptomycin sulphate (2 mg/ml) + Penicillin G (1,500 U/ml)	Oral	22 days Once daily	Simultaneous	Forced Swim Test, Corticosterone, MDA, Body Weight, Fecal Boli, Blood LPS & LBP, mRNA Relative Expression of TLR-4, MD-2, NF-kappa B p65, I-kappa-B-alpha, COX-2, & IL-1 beta, Protein Expression TLR-4, MD-2, & NF-kappa B p65,

									Levels of prostaglandin PGE2 & 15d-PGJ2, NF- B p65 activity
Liang et al.,	2015	Chronic Restraint Stress	L. helveticus NS8	Probiotic	1.00E+09 (CFU)	Oral	26 days Once daily	Simultaneous	Object Recognition, Sucrose Preference Test, BDNF, Serotonin, Dopamine, Norepinephrine, Corticosterone, ACTH Body Weight, IL-10, IFN-Y, TNF-alpha
Lui et al.,	2016	Germ-free Housing	L. plantarum PS128	Probiotic	5E+09 (CFU)	Oral	16 days Once daily	Pre	Forced Swim Test, Dopamine, DOPAC, HVA, Serotonin, 5-HIAA Corticosterone, Body Weight, Cecum Weight
Savignac et al.,	2014	BALB/cOla Hsd	B. Longum 1714	Probiotic	1E+09 (CFU)	Oral	42 days Once daily	Post	Forced Swim Test, Tail Suspension Test, Corticosterone, Body Weight
		BALB/cOla Hsd	B. Breve 1205	Probiotic	1E+09 (CFU)	Oral			

Wong et al.,	2016	Chronic Restraint Stress	Minocycline	Antibiotic	5 (mg/kg)	IP	21 days Once daily	Simultaneous	Forced Swim Test, Respirometry, Relative abundance of: Turicibacter, Bifidobacterium, Akkermansia, Blautia, Lachnospiraceae, & Lactobacillus
Zheng et al.,	2016	Germ-free Housing + MDD Patient Fecal Sample	Healthy Fecal Matter	Other	0.1 (g)	Coloni-zation	Once	Post	Forced Swim Test, Tail Suspension Test

Table 2. Study Design Characteristics of Gut Microbiota-targeting Interventions to Induce Depression

Authors	Year	Species	Strain	Sex	Model	Model Category	Outcomes Measured
Ait-Belgnaoi et al.,	2014	Mouse	C57BL/6	Male	Water Avoidance Stress	Indirect	FOS-Immunoreactivity, Corticosterone, Adrenaline, Noradrenaline, Occulin & JAMA-A Protein Expression, Doublecortin- expressing cells, % 51-Cre-EDTA recovery
Bravo et al.,	2011	Mouse	Balb/c	Male	Forced Swim Test	Indirect	Corticosterone
Desbonnet et al.,	2010	Rat	Sprague Dawley	Male	Maternal Separation	Indirect	Forced Swim Test, Corticosterone, Noradrenaline, Dopamine, Serotonin, DOPAC, 5HIAA, HVA, Tryptophan, L-Kynurenine, Kynurenic acid, IL6 (LPS), IFN-Y (LPS), TNF-alpha (LPS), IL-10 (LPS), Body Weight
Farshim et al.,	2016	Rat	Wistar	Male	Prolonged Weaning	Direct	Oxytocin Receptor Binding, Forced Swim Test, Total Bacteria, Clostridium histolyticum, Lactobacillus-Enterococcus, Bifidobacterium spp.
Gacias et al.,	2016	Mouse	C57BL/6	Male	Gastric Gavage	Direct	Forced Swim Test, Social Interaction
		Mouse	NOD	Male	Gastric Gavage	Direct	
Gárate et al.,	2011	Rat	Sprague Dawley	Male	Chronic Mild Stress	Indirect	Forced Swim Test, Corticosterone, MDA, Body Weight, Fecal Boli, Blood LPS & LBP, mRNA Relative Expression of TLR-4, MD-2, NF-kappa B p65,

							I-kappa-B-alpha, COX-2, & IL-1 beta, Protein Expression TLR-4, MD-2, & NF-kappa B p65, Levels of prostaglandin PGE2 & 15d-PGJ2, NF- B p65 activity
Ilgin et al.,	2015	Rat	Wistar	Female	Ciprofloxacin (CPX)	Direct	Noradrenaline, Adrenaline, Dopamine, Serotonin, GABA, Glutamate, Forced Swim Test, MDA, Glutathione, Superoxide dismutase, Catalase
Jørgensen et al.,	2014	Mouse	BALB/cAnNTac	Male	High Fat Diet	Direct	Forced Swim Test, Morris Water Maze, Body Weight
		Mouse	BALB/cAnNTac	Male	High Sucrose Diet	Direct	
Liang et al.,	2015	Rat	Sprague Dawley	Male	Chronic Restraint Stress	Indirect	Object Recognition, Sucrose Preference Test, BDNF, Serotonin, Dopamine, Norepinephrine, Corticosterone, ACTH Body Weight, IL-10, IFN-Y, TNF-alpha
O'Mahony et al.,	2009	Mouse	Sprague Dawley	Not Reported	Maternal Separation	Indirect	Corticosterone, Fecal Boli, TNF-alpha (LPS), IL-6 (LPS), IFN-Y (LPS), IL-4 (LPS), IL-10 (LPS)
Winther et al.,	2015	Mouse	C57BL/6	Male	Magnesium Deficient Diet	Direct	Forced Swim Test, Body Weight
Wong et al.,	2016	Mouse	C57BL/6J	Male	Chronic Restraint Stress	Indirect	Respirometry, Relative abundance of: Turicibacter, Bifidobacterium, Akkermansia, Blautia, Lachnospiraceae, & Lactobacillus
Zheng et al.,	2016	Mouse	Kunming	Male	Germ-free Housing	Direct	Forced Swim Test

Table 3. Reporting of Measures to Reduce the Risk of Bias in Gut Microbiota-targeting interventions to Reduce Depression

Authors	**Year**	Randomise to Treatment	Allocation Concealment	Blinded Outcome Assessment	Sample Size Calculation	Conflict of Interest Statement	Compliance with Welfare Regulations	Source of Funding
Ait-Belgnaoi et al.,	2014	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	Private
Bravo et al.,	2011	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	Public
Desbonnet et al.,	2010	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	Public
Gacias et al.,	2016	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	Both
Gárate et al.,	2011	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	Both
Liang et al.,	2015	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	Private
Lui et al.,	2016	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	Public
Savignac et al.,	2014	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	Public
Wong et al.,	2016	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	Public
Zheng et al.,	2016	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	Public
	Total	4/10=40%	0/13 = 0%	4/10 = 40%	1/10=10%	9/10=90%	100%	

Table 4. Reporting of Measures to Reduce the Risk of Bias in Gut Microbiota-targeting Interventions to Induce Depression

Authors	**Year**	Randomise to Model/Control	Allocation Concealment	Blinded Outcome Assessment	Sample Size Calculation	Conflict of Interest Statement	Compliance with Welfare Regulations	Source of Funding
Ait-Belgnaoi et al.,	2014	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	Private
Bravo et al.,	2011	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	Both
Desbonnet et al.,	2010	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	Public
Farshim et al.,	2016	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	Public
Gacias et al.,	2016	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	Both
Gárate et al.,	2011	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	Both
Ilgin et al.,	2015	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	Not Reported
Jørgensen et al.,	2014	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	None
Liang et al.,	2015	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	Private
O'Mahony et al.,	2009	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	Both
Winther et al.,	2015	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	Public
Wong et al.,	2016	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	Public
Zheng et al.,	2016	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	Public
	Total	4/13 = 30.8%	0/13 0%	7/13 = 53.8%	1/13 = 7.6%	12/13 = 92.3%	12/13 = 92.3%	

APPENDIX 4: Summary Tables from Chapter 6: THE ANTIDEPRESSANT EFFECT OF KETAMINE IN ANIMAL MODELS OF DEPRESSION AS MEASURED BY THE FORCED SWIM TEST: A SYSTEMATIC REVIEW AND META-ANALYSIS

Table 1. Study Design Characteristics of experiments included in the ketamine systematic review.

Surname	Year	Disease Model	Timing of Admin	Freq	Ketamine Type	Dose	Sex	Species	Time
Akinfiresoye et al.	2013	Genetic	During	>1	Ketamine	0.25	Male	Rat	22
Akinfiresoye et al.	2013	Genetic	During	>1	Ketamine	0.5	Male	Rat	22
Antony et al.	2014	Forced Swim Test	Before	1	Ketamine	5	Male	Mouse	0.33
Antony et al.	2014	Forced Swim Test	Before	1	Ketamine	7.5	Male	Mouse	0.33
Antony et al.	2014	Forced Swim Test	Before	1	Ketamine	10	Male	Mouse	0.33
Assis et al.	2009	Forced Swim Test	After	1	Ketamine	5	Male	Rat	1
Assis et al.	2009	Forced Swim Test	After	1	Ketamine	10	Male	Rat	1
Assis et al.	2009	Forced Swim Test	After	1	Ketamine	15	Male	Rat	1
Autry et al.	2011	Forced Swim Test	Before	1	Ketamine	3	Male	Mouse	0.5
Autry et al.	2011	Forced Swim Test	Before	1	Ketamine	3	Male	Mouse	3
Autry et al.	2011	Forced Swim Test	Before	1	Ketamine	3	Male	Mouse	24
Autry et al.	2011	Forced Swim Test	Before	1	Ketamine	3	Male	Mouse	168
Baptista et al.	2015	Chronic Stress	After	1	Ketamine	10	Male	Rat	16
Burgdorf et al.	2013	Forced Swim Test	After	1	Ketamine	10	Male	Rat	24
Carrier et al.	2013	Forced Swim Test	Before	1	Ketamine	2.5	Male	Rat	0.5
Carrier et al.	2013	Forced Swim Test	Before	1	Ketamine	2.5	Female	Rat	0.5
Carrier et al.	2013	Forced Swim Test	Before	1	Ketamine	5	Female	Rat	0.5
Carrier et al.	2013	Forced Swim Test	Before	1	Ketamine	5	Male	Rat	0.5
Carrier et al.	2013	Forced Swim Test	Before	1	Ketamine	10	Male	Rat	0.5
Carrier et al.	2013	Forced Swim Test	Before	1	Ketamine	10	Female	Rat	0.5
Chaturvedi et al.	1999	Learned helplessness	Before	1	Ketamine	2.5	Both	Mouse	0.5

Chaturvedi et al.	2001	Chronic Stress	Before	1	Ketamine	2.5	Both	Mouse	26
Chaturvedi et al.	1999	Learned helplessness	Before	1	Ketamine	5	Both	Mouse	0.5
Chaturvedi et al.	2001	Chronic Stress	Before	1	Ketamine	5	Both	Mouse	26
Chaturvedi et al.	1999	Learned helplessness	Before	1	Ketamine	10	Both	Mouse	0.5
Chaturvedi et al.	2001	Chronic Stress	Before	1	Ketamine	10	Both	Mouse	26
Chiu et al.	2015	Chronic Stress	Before	1	Ketamine	2.5	Male	Mouse	1.5
Chiu et al.	2015	Chronic Stress	Before	1	Ketamine	25	Male	Mouse	1.5
Chiu et al.	2015	Chronic Stress	Before	1	Ketamine	50	Male	Mouse	1.5
Chiu et al.	2015	Chronic Stress	Before	1	Ketamine	50	Male	Mouse	168
Chiu et al.	2015	Chronic Stress	Before	1	Ketamine	50	Male	Mouse	336
Cruz et al.	2009	Forced Swim Test	Before	1	Ketamine	6.25	Male	Mouse	0.5
Cruz et al.	2009	Forced Swim Test	Before	1	Ketamine	12.5	Male	Mouse	0.5
Cruz et al.	2009	Forced Swim Test	Before	1	Ketamine	25	Male	Mouse	0.5
Cruz et al.	2009	Forced Swim Test	Before	1	Ketamine	50	Male	Mouse	0.5
da Silva et al.	2010	Forced Swim Test	Before	1	Ketamine	5	Male	Mouse	0.17
da Silva et al.	2010	Forced Swim Test	Before	1	Ketamine	10	Male	Mouse	0.17
da Silva et al.	2010	Forced Swim Test	Before	1	Ketamine	20	Male	Mouse	0.17
Engin et al.	2009	Forced Swim Test	After	1	Ketamine	10	Male	Rat	0.5
Engin et al.	2009	Forced Swim Test	After	1	Ketamine	50	Male	Rat	0.5
Franceschelli et al.	2015	Chronic Stress	After	1	Ketamine	10	Female	Mouse	0.5
Franceschelli et al.	2015	Chronic Stress	After	1	Ketamine	10	Male	Mouse	0.5
Franceschelli et al.	2015	Chronic Stress	After	1	Ketamine	10	Male	Mouse	168
Franceschelli et al.	2015	Chronic Stress	After	1	Ketamine	10	Female	Mouse	168
Fuchikami et al.	2015	Forced Swim Test	After	1	Ketamine	3	Male	Rat	24
Fuchikami et al.	2015	Forced Swim Test	After	1	Ketamine	10	Male	Rat	24
Fuchikami et al.	2015	Forced Swim Test	After	1	Ketamine	30	Male	Rat	24
Garcia et al.	2008	Forced Swim Test	After	1	Ketamine	5	Male	Rat	1

Garcia et al.	2008	Forced Swim Test	Before	>1	Ketamine	5	Male	Rat	1
Garcia et al.	2008	Forced Swim Test	After	1	Ketamine	10	Male	Rat	1
Garcia et al.	2008	Forced Swim Test	Before	>1	Ketamine	10	Male	Rat	1
Garcia et al.	2008	Forced Swim Test	Before	>1	Ketamine	15	Male	Rat	1
Garcia et al.	2008	Forced Swim Test	After	1	Ketamine	15	Male	Rat	1
Ghasemi et al.	2010	Forced Swim Test	Before	1	Ketamine	0.5	Male	Mouse	0.75
Ghasemi et al.	2010	Forced Swim Test	Before	1	Ketamine	1	Male	Mouse	0.75
Ghasemi et al.	2010	Forced Swim Test	Before	1	Ketamine	2	Male	Mouse	0.75
Ghasemi et al.	2010	Forced Swim Test	Before	1	Ketamine	5	Male	Mouse	0.75
Gideons et al.	2014	Forced Swim Test	Before	1	Ketamine	3	Male	Mouse	0.5
Gideons et al.	2014	Forced Swim Test	Before	1	Ketamine	3	Male	Mouse	8
Gideons et al.	2014	Forced Swim Test	Before	1	Ketamine	3	Male	Mouse	24
Gigliucci et al.	2013	Chronic Stress	After	1	Ketamine	25	Male	Rat	24
Haj-Mirzaian et al.	2015	Chronic Stress	After	1	Ketamine	1	Male	Mouse	1
Jett et al.	2015	Forced Swim Test	Before	1	Ketamine	10	Male	Rat	168
Kilic et al.	2014	Forced Swim Test	Before	1	Ketamine	10	Female	Rat	
Koike et al.	2013	CORT/DEX/LPS/ACTH Insult	After	>1	Ketalar	1	Male	Rat	0.5
Koike et al.	2014	Forced Swim Test	After	1	Ketalar	1	Male	Rat	24
Koike et al.	2013	CORT/DEX/LPS/ACTH Insult	After	>1	Ketalar	3	Male	Rat	0.5
Koike et al.	2014	Forced Swim Test	After	1	Ketalar	3	Male	Rat	24
Koike et al.	2013	CORT/DEX/LPS/ACTH Insult	After	>1	Ketalar	10	Male	Rat	0.5
Koike et al.	2014	Forced Swim Test	After	1	Ketalar	10	Male	Rat	24
Li et al.	2015	Forced Swim Test	Before	1	Ketamine	3	Male	Mouse	0.5
Li et al.	2015	Forced Swim Test	Before	1	Ketamine	10	Male	Mouse	0.5
Li et al.	2014	CORT/DEX/LPS/ACTH Insult	After	1	Ketamine	10	Both	Mouse	46
Li et al.	2015	Forced Swim Test	Before	1	Ketamine	20	Male	Mouse	0.5
Liebenberg et al.	2015	Forced Swim Test	During	1	Ketamine	15	Male	Rat	1

Ma et al.	2013	Chronic Stress	During	1	Ketamine	10	Male	Mouse	3
Ma et al.	2013	Chronic Stress	During	1	Ketamine	10	Male	Mouse	24
Ma et al.	2013	Chronic Stress	Before	1	Ketamine	10	Male	Mouse	48
Mlyniec et al.	2015	Forced Swim Test	After	>1	Ketamine	0.25	Male	Mouse	24
Moreira et al.	2016	Chronic Stress	After	1	Ketamine	10	Female	Rat	0.5
Nosyreva et al.	2014	Forced Swim Test	Before	1	Ketamine	3	Male	Mouse	24
Parise et al.	2013	Forced Swim Test	After	>1	Ketamine	5	Male	Rat	24
Parise et al.	2013	Forced Swim Test	After	>1	Ketamine	10	Male	Rat	24
Parise et al.	2013	Chronic Stress	After	1	Ketamine	20	Male	Rat	1
Parise et al.	2013	Forced Swim Test	After	>1	Ketamine	20	Male	Rat	24
Perrine et al.	2014	Chronic Stress	After	1	Ketamine	40	Male	Rat	21
Petryshen et al.	2016	Forced Swim Test	After	1	Ketamine	3	Male	Mouse	0.5
Petryshen et al.	2016	Forced Swim Test	After	1	Ketamine	10	Male	Mouse	0.5
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	1.25	Male	Mouse	0.5
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	1.25	Male	Mouse	336
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	2.5	Male	Mouse	0.5
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	2.5	Male	Mouse	336
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	5	Male	Mouse	0.5
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	5	Male	Mouse	336
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	10	Male	Mouse	0.5
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	10	Male	Mouse	336
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	50	Male	Mouse	0.5
Popik et al.	2008	Forced Swim Test	During	>1	Ketamine	50	Male	Rat	0.67
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	50	Male	Mouse	336
Popik et al.	2008	Forced Swim Test	During	1	Ketamine	160	Male	Rat	168
Pozzi et al.	2014	Forced Swim Test	Before	1	Ketamine	3	NR	Mouse	0.5
Pozzi et al.	2014	Forced Swim Test	Before	1	Ketamine	3	NR	Mouse	24

Pozzi et al.	2014	Forced Swim Test	Before	1	Ketamine	3	NR	Mouse	168
Prabhakar et al.	2011	Forced Swim Test	Before	1	Ketamine	2	Male	Mouse	0.5
Ren et al.	2016	Genetic	Before	1	Ketamine	3	Female	Mouse	8
Ren et al.	2015	Other	After	1	Ketamine	10	Male	Mouse	1
Reus et al.	2011	Forced Swim Test	After	1	Ketamine	5	Male	Rat	1
Reus et al.	2015	Chronic Stress	After	1	(S)-ketamine	15	Male	Rat	0
Reus et al.	2014	Forced Swim Test	During	>1	Ketamine	15	NR	Rat	1
Reus et al.	2015	Chronic Stress	After	>1	Ketamine	15	Male	Rat	1
Reus et al.	2013	Chronic Stress	Before	>1	Ketamine	15	Male	Rat	24
Salat et al.	2015	Forced Swim Test	Before	1	NorKetamine	5	Male	Mouse	0.5
Salat et al.	2015	Forced Swim Test	Before	1	Ketamine	5	Male	Mouse	0.5
Salat et al.	2015	Forced Swim Test	Before	1	DehydroNorKetamine	5	Male	Mouse	0.5
Salat et al.	2015	Forced Swim Test	Before	1	NorKetamine	10	Male	Mouse	0.5
Salat et al.	2015	Forced Swim Test	Before	1	DehydroNorKetamine	10	Male	Mouse	0.5
Salat et al.	2015	Forced Swim Test	Before	1	Ketamine	10	Male	Mouse	0.5
Salat et al.	2015	Forced Swim Test	Before	1	DehydroNorKetamine	50	Male	Mouse	0.5
Salat et al.	2015	Forced Swim Test	Before	1	NorKetamine	50	Male	Mouse	0.5
Salat et al.	2015	Forced Swim Test	Before	1	Ketamine	50	Male	Mouse	0.5
Salat et al.	2015	Forced Swim Test	Before	1	Ketamine	50	Male	Mouse	72
Salat et al.	2015	Forced Swim Test	Before	1	NorKetamine	50	Male	Mouse	72
Salat et al.	2015	Forced Swim Test	Before	1	Ketamine	50	Male	Mouse	168
Salat et al.	2015	Forced Swim Test	Before	1	NorKetamine	50	Male	Mouse	168
Sarkar et al.	2016	Chronic Stress	After	1	Ketamine	2.5	Male	Rat	27
Sarkar et al.	2016	Chronic Stress	After	1	Ketamine	2.5	Female	Rat	27
Sarkar et al.	2016	Chronic Stress	After	1	Ketamine	5	Female	Rat	27

Sarkar et al.	2016	Chronic Stress	After	1	Ketamine	5	Male	Rat	27
Sun et al.	2016	Chronic Stress	After	1	Ketamine	10	Male	Rat	0.5
Sun et al.	2016	Chronic Stress	After	1	Ketamine	10	Male	Rat	72
Tang et al.	2015	Chronic Stress	After	1	Ketamine	30	Male	Mouse	24
Tang et al.	2015	Chronic Stress	After	1	Ketamine	30	Male	Mouse	120
Tizabi et al.	2012	Genetic	Before	1	Ketamine	0.5	Female	Rat	0.5
Tizabi et al.	2012	Genetic	Before	>1	Ketamine	0.5	Female	Rat	22
Tizabi et al.	2012	Genetic	Before	>1	Ketamine	0.5	Female	Rat	168
Tizabi et al.	2012	Genetic	Before	1	Ketamine	2.5	Female	Rat	0.5
Tizabi et al.	2012	Genetic	Before	>1	Ketamine	2.5	Female	Rat	22
Tizabi et al.	2012	Genetic	Before	1	Ketamine	2.5	Female	Rat	168
Tizabi et al.	2012	Genetic	Before	>1	Ketamine	2.5	Female	Rat	168
Tizabi et al.	2012	Genetic	Before	>1	Ketamine	2.5	Female	Rat	336
Tizabi et al.	2012	Genetic	Before	1	Ketamine	5	Female	Rat	168
Vogt et al.	2015	Forced Swim Test	Before	>1	Ketamine	30	Male	Mouse	1
Vogt et al.	2015	Forced Swim Test	Before	>1	Ketamine	30	Male	Mouse	24
Walker et al.	2013	CORT/DEX/LPS/ACTH Insult	After	1	Ketamine	6	Male	Mouse	28
Walker et al.	2015	CORT/DEX/LPS/ACTH Insult	After	1	Ketamine	10	Male	Rat	24
Wang et al.	2015	Chronic Stress	After	1	Ketamine	10	Male	Rat	0.5
Wang et al.	2015	Chronic Stress	After	1	Ketamine	10	Male	Rat	1
Wang et al.	2015	Chronic Stress	After	1	Ketamine	10	Male	Rat	2
Wang et al.	2015	Chronic Stress	After	1	Ketamine	10	Male	Rat	4
Wang et al.	2011	Other	After	1	Ketamine	10	Male	Rat	24
Wrobel et al.	2015	CORT/DEX/LPS/ACTH Insult	After	>1	Ketamine	15	Male	Mouse	0.5
Wrobel et al.	2015	CORT/DEX/LPS/ACTH Insult	After	1	Ketamine	30	Male	Mouse	0.5
Xia et al.	2016	Learned helplessness	After	1	Ketamine	30	NR	Mouse	0.5
Xia et al.	2016	Learned helplessness	After	1	Ketamine	30	NR	Mouse	120

Xu et al.	2013	Forced Swim Test	After	1	Ketamine	10	Male	Rat	0.5
Yang et al.	2013	Forced Swim Test	After	1	Ketamine	5	Male	Rat	0.5
Yang et al.	2013	Forced Swim Test	After	1	Ketamine	10	Male	Rat	0.5
Yang et al.	2013	CORT/DEX/LPS/ACTH Insult	After	1	Ketamine	10	Male	Rat	1
Yang et al.	2015	Chronic Stress	After	1	(S)-ketamine	10	Male	Mouse	48
Yang et al.	2015	Chronic Stress	After	1	(R)-ketamine	10	Male	Mouse	48
Yang et al.	2015	Chronic Stress	After	1	(S)-ketamine	10	Male	Mouse	168
Yang et al.	2015	Chronic Stress	After	1	(R)-ketamine	10	Male	Mouse	168
Yang et al.	2013	Forced Swim Test	After	1	Ketamine	15	Male	Rat	0.5
Yilmaz et al.	2002	Forced Swim Test	After	1	Ketamine	160	Male	Rat	96
Yilmaz et al.	2002	Forced Swim Test	After	1	Ketamine	160	Male	Rat	192
Yilmaz et al.	2002	Forced Swim Test	After	1	Ketamine	160	Male	Rat	264
Zanos et al.	2015	Forced Swim Test	After	1	Ketamine	10	Male	Mouse	1
Zanos et al.	2015	Forced Swim Test	After	1	Ketamine	10	Male	Mouse	24
Zhang et al.	2013	Forced Swim Test	After	1	Ketamine	10	Male	Rat	1
Zhang et al.	2015	Chronic Stress	After	1	Ketamine	10	Male	Mouse	5
Zhang et al.	2015	Chronic Stress	Before	>1	Ketamine	10	Male	Rat	24
Zhang et al.	2014	CORT/DEX/LPS/ACTH Insult	After	1	(S)-ketamine	10	Male	Mouse	24
Zhang et al.	2014	CORT/DEX/LPS/ACTH Insult	After	1	(R)-ketamine	10	Male	Mouse	24
Zhang et al.	2015	Chronic Stress	Before	>1	Ketamine	10	Male	Rat	72
Zhang et al.	2015	Chronic Stress	Before	>1	Ketamine	10	Male	Rat	120
Zhang et al.	2015	Chronic Stress	Before	>1	Ketamine	10	Male	Rat	168
Zhang et al.	2014	CORT/DEX/LPS/ACTH Insult	After	1	(S)-ketamine	10	Male	Mouse	168
Zhang et al.	2014	CORT/DEX/LPS/ACTH Insult	After	1	(R)-ketamine	10	Male	Mouse	168
Zhang et al.	2015	Chronic Stress	Before	>1	Ketamine	10	Male	Rat	192
Zhang et al.	2015	Chronic Stress	Before	>1	Ketamine	10	Male	Rat	288
Zhang et al.	2015	Chronic Stress	Before	>1	Ketamine	10	Male	Rat	360

Zhang et al.	2015	Chronic Stress	Before	>1	Ketamine	10	Male	Rat	456

Surname = the first author's surname, Year = year the study was published, Disease Model = the method of disease model induction, Timing of Admin = whether ketamine was administered before, during, or after model induction, Freq = the frequency of ketamine administered (once or more than once), Ketamine Type = the form of ketamine administered, Dose = the dose of ketamine administered (mg/kg), Sex = sex of the animals used, Species, Time = the time (in mins) ketamine was administered prior to outcome assessment, NR = Not Reported.

Table 2. Reporting of Measures to Reduce the Risk of Bias in studies investigating ketamine

Author	Year	Randomisation	Blinding	Allocation	SSC	CoI	Welfare Regs.	Source of Funding
Akinfiresoye et al.	2013	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Antony et al.	2014	Yes	Yes	Yes	0	Yes	Yes	Unknown
Assis et al.	2009	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Public
Autry et al.	2011	Not Reported	Not Reported	Not Reported	0	No	Yes	Unknown
Baptista et al.	2015	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Burgdorf et al.	2013	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Unknown
Carrier et al.	2013	Not Reported	Yes	Not Reported	0	Not Reported	Yes	Unknown
Chaturvedi et al.	2001	Not Reported	Not Reported	Not Reported	0	No	Not Reported	Public
Chaturvedi et al.	1999	Yes	Not Reported	Not Reported	0	Not Reported	Yes	Public
Chiu et al.	2015	No	Not Reported	No	0	Yes	Yes	Public
Cruz et al.	2009	Not Reported	Yes	Not Reported	0	Not Reported	Yes	Public
da Silva et al.	2010	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Unknown
Engin et al.	2009	Yes	Not Reported	Not Reported	0	No	Yes	Public
Franceschelli et al.	2015	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Both
Fuchikami et al.	2015	Yes	Yes	Not Reported	0	Not Reported	Yes	Both
Garcia et al.	2008	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Public
Garcia et al.	2008	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Public
Ghasemi et al.	2010	Not Reported	Not Reported	Not Reported	0	No	Yes	Public

Gideons et al.	2014	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Gigliucci et al.	2013	Not Reported	Yes	Not Reported	0	Not Reported	Yes	Public
Haj-Mirzaian et al.	2015	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Unknown
Jett et al.	2015	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Kilic et al.	2014	Not Reported	Yes	Not Reported	0	Yes	Yes	Unknown
Koike et al.	2014	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Unknown
Koike et al.	2013	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Unknown
Li et al.	2015	Not Reported	Yes	Not Reported	0	Not Reported	Yes	Public
Li et al.	2014	Not Reported	Yes	Not Reported	0	Not Reported	Yes	Public
Liebenberg et al.	2015	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Ma et al.	2013	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Mlyniec et al.	2015	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Moreira et al.	2016	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Nosyreva et al.	2014	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Parise et al.	2013	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Perrine et al.	2014	Not Reported	Yes	Not Reported	0	Not Reported	Yes	Public
Petryshen et al.	2016	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Private
Popik et al.	2008	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Pozzi et al.	2014	Yes	Not Reported	Not Reported	0	Yes	Yes	Public
Prabhakar et al.	2011	Yes	Yes	Not Reported	0	Not Reported	Yes	Unknown
Ren et al.	2016	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public

Ren et al.	2015	No	Not Reported	Not Reported	0	Yes	Yes	Both
Reus et al.	2014	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Reus et al.	2011	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Reus et al.	2015	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Public
Reus et al.	2013	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Public
Reus et al.	2015	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Public
Salat et al.	2015	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Sarkar et al.	2016	Not Reported	Not Reported	Not Reported	0	No	Yes	Public
Sun et al.	2016	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Tang et al.	2015	Not Reported	Yes	Not Reported	0	Not Reported	Yes	Unknown
Tizabi et al.	2012	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Public
Vogt et al.	2015	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Walker et al.	2013	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Walker et al.	2015	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Wang et al.	2015	Yes	Yes	Not Reported	0	Yes	Yes	Public
Wang et al.	2011	Yes	Yes	Not Reported	0	No	Yes	Unknown
Wrobel et al.	2015	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Xia et al.	2016	Not Reported	Yes	Not Reported	0	Yes	Yes	Public
Xu et al.	2013	Yes	Not Reported	Not Reported	0	Yes	Yes	Public
Yang et al.	2013	Yes	Not Reported	Not Reported	0	Yes	Yes	Public
Yang et al.	2015	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Both

Yang et al.	2013	Yes	Not Reported	Not Reported	0	Not Reported	Yes	Public
Yilmaz et al.	2002	Not Reported	Not Reported	Not Reported	0	No	Yes	Public
Zanos et al.	2015	Not Reported	Yes	Not Reported	0	Not Reported	Yes	Public
Zhang et al.	2013	Yes	Not Reported	Not Reported	0	No	Yes	Public
Zhang et al.	2015	Not Reported	Not Reported	Not Reported	0	Yes	Yes	Public
Zhang et al.	2014	Not Reported	Not Reported	Not Reported	0	No	Yes	Public
Zhang et al.	2015	Not Reported	Not Reported	Not Reported	0	Not Reported	Yes	Public
Percentage Reporting Compliance		12/67 = 17.9%	25/67 = 37.4%	2/67 = 2.9%	0%	32/67 = 47.8%	66/67 = 98.5%	55/67 = 82.1%

Randomisation = Random allocation to group, Blinding = Blinded assessment of outcome, Allocation = Allocation concealment, CoI = Conflict of interest, SSC = Sample size calculation, Welfare Regs. = Compliance with animal welfare regulations, Source of Funding = Source of funding. Percentage reporting compliance = the number of papers that reported the measure to reduce the risk of bias.

APPENDIX 5: R CODE TO CALCULATE A PRIOR SAMPLE SIZE CALCULATION

```
## means and standard deviations extracted from graphs
## in McVey Neufeld et al., 2017 (Mc), and Burokas et al., 2017 (Burokas)



## sum of squares from ANOVA tables from the above means and standard
deviations


SSeffectMc <- 529.7
SStotalMc <- 1014

ssEffectBurokas <- 32486.91
ssTotalBurokas <- 56028.04



## calculating eta squared
etasqrdMc <- SSeffectMc/SStotalMc
etasqrdMc

etasqrdBurokas <- ssEffectBurokas / ssTotalBurokas
etasqrdBurokas



## calculating effect size f
fEffectMc <- sqrt(etasqrdMc/(1-etasqrdMc))
fEffectMc

fEffectBurokas <- sqrt(etasqrdBurokas/(1-etasqrdBurokas))
fEffectBurokas

## install packages for power calcs
install.packages("pwr", "pwr2")
library(pwr, pwr2)
```

```r
# effect size f is calculated by
# between group standard deviation (SD of k means) / within group standard deviation (SD of k groups)

pwr.anova.test(k = 4, n = NULL, f = fEffectMc, sig.level = 0.01, power = 0.9)

pwr.anova.test(k = 4, n = NULL, f = fEffectBurokas, sig.level = 0.01, power = 0.9)



# to see what our study will be powered at with 8 animals per group
pwr.anova.test(k = 4, n = 8, f = fEffectMc, sig.level = 0.01, power = NULL)

# to see what effect we can estimate with the n and power
pwr.anova.test(k = 4, n = 8, f = NULL, sig.level = 0.01, power = 0.9)
```