Patent Landscape Reporting Quality and Predicting Drug Approval

James Andrew Smith

Kellogg College

University of Oxford

Thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2018

Abstract

Attrition is a major contributor to the cost of new drug development. Therefore, the ability to predict drug approval could improve R&D efficiency. The success of drug development programs is governed both by scientific factors and by “external” factors, such as intellectual property, company characteristics, and economics. Incorporating these external factors into predictions of drug approval could help to improve decision-making.

We postulated that patents might represent a particularly under-used information source in this context. However, in reviewing the academic patent literature, it was apparent that improvements in the quality of available evidence were needed before patent landscapes, a common form of patent analysis, could be useful. To examine the extent and severity of this problem, we conducted a systematic review of the reporting quality of patent landscape articles. Finding evidence that reporting was insufficient, we developed the Reporting Items for Patent Landscapes (RIPL) statement, a reporting guideline which aims to improve it.

Subsequently, we focussed on predicting approval and failure of drug candidates. A systematic review identified only three papers developing multivariable models to predict approval on the basis of analysis of approved and failed drugs. These models are of limited utility due to methodological and reporting quality issues. Therefore, we developed and internally validated two models to predict approval: one incorporating candidate predictor variables to represent the external factors previously mentioned, as well as physicochemical parameters, and the second including only physicochemical parameters. Performance was modest but positive and further work is needed to externally validate the models.

Acknowledgements

I would like to thank my supervisors, Prof Andy Carr and Dr David Brindley, for providing intellectual support and for giving me freedom to explore the research questions that I found interesting. I am grateful to all of the members of both David's and Andy's groups who have provided input on this thesis. In particular, Céline Halioua has provided a great deal of support and Dr Michelle Van Velthoven has given useful feedback and contributed to the research. Thanks are due to both Alison Carter and Ann Watson for their administrative support.

I am extremely grateful for Medical Research Council UK Funding for this DPhil. I would like to thank the Centre for Advancement of Sustainable Medical Innovation (CASMI) Translational Stem Cell Consortium (CTSCC), particularly Kim Bure and David, and the SENS Research Foundation (SRF), including the SRF-Oxford-CTSCC Research Program. Both organisations led to my interest in translational research and since then have introduced me to many people whose support has been essential. Of those, I would in particular like to acknowledge Zeeshaan Arshad for his contributions to the research in this thesis. In Chapter 2 he conducted independent data extraction from the identified papers and assisted in the statistical analysis, and generally helped to plan the research. In Chapter 3, he helped with administering, designing, and analysing the modified Delphi study and screening participants for eligibility. I would also like to thank him for reviewing Chapter 4, and for his input on pilot studies of the research that resulted in Chapter 5. Thank you to Felix Jackson and Sumedh Sontakke for their work as Summer Scholars on the same pilot studies, and Prof Chas Bountra and Dr Wen Hwa Lee for providing mentorship and input on those studies. Thanks are due to Zain Hussain and Michelle Van Velthoven for providing an independent review of titles, abstracts, and full texts for Chapter 4 and Piotr Pawlik for writing the Python script to collect literature and patent counts for compounds in Chapter 5.

I owe special thanks to Dr Paula Dhiman for her indispensable input on the methodology for Chapter 5 and review of aspects of Chapters 2 and 4. I would also like to thank Prof Gary Collins for his invaluable input into the methodology for Chapter 5 during the Confirmation of Status and in a later meeting, and for introducing me to Paula. Prof Peter Taylor provided very useful comments during the Confirmation of Status which stimulated much of the discussion relating to Chapter 5, and for which I am grateful. I would also like to thank Dr Tom Kaiser for taking the time to show me how to use KNIME and for being an incredibly useful source of knowledge for any chemistry-related queries.

I would like to thank David for providing access to the Cortellis data used in Chapter 5 and Dr Liliane Zorzela for providing data from her systematic review of adverse event reporting which was, in the end, not included in this thesis. I would also like to extend my gratitude to all of the participants in the Delphi study reported in Chapter 3 and to the EQUATOR Network for their support in developing the guideline. Dr Iveta Simera, in particular, provided useful input in the early stages of planning Chapter 2. Thank you to Kayla Friedman and Malcolm Morgan of the University of Cambridge and Charles Weir of Lancaster University for producing the Microsoft Word thesis template adapted for use with this thesis.

Major components of Chapter 2 first appeared in the following paper: “Smith JA, Arshad Z, Thomas H, Carr AJ, Brindley DA: Evidence of insufficient quality of reporting in patent landscapes in the life sciences. Nature Biotechnology, 210-214, 2017” by Springer Nature. Major components of Chapter 3 have been accepted for publication in the November issue in 2018 of Nature Biotechnology, by Springer Nature. For Chapter 4, a protocol has been published on PROSPERO. Details are provided in each chapter and all co-authors have agreed for the material to appear in this thesis.

Many thanks to James Walker and Dad for suffering through proof-reading most of the thesis, and Ben Tarring for doing the same for the discussion.

Finally, thanks to Mum, Dad and Rory for providing tea and distractions when needed during the write-up.

Contents

List of Tables

List of Figures

1 Introduction

Biomedical research has resulted in considerable advancement of human health. People live longer than ever before, there are more treatments for more indications, and there is an exponentially growing knowledge base from which to develop new medicines or otherwise improve health. Despite this, there is evidence suggesting that the efficiency of biomedical research has been consistently falling over the last 50 years. Health outcomes have not kept pace with research outputs[1], the cost per new molecular entity (NME) brought to market has increased dramatically[2,3], the financial performance of the biotechnology sector is poor and declining[4], and the sustainability of biomedical innovation has been called into question[5]. There is widespread concern over the reproducibility of[6], as well as the truth of claims in[7], much published research literature. Here, we aim to identify contributors to this inefficiency and develop approaches to improve it.

Reducing attrition in drug development is a natural starting point for efforts to reduce wastage. Recent estimates indicate success rates of 13.8% for NMEs first entering man[8], and the cost of drug development is therefore dominated by the cost of failures. The ability to better identify and prioritise those compounds which are eventually approved would therefore be of considerable value. A huge body of research exists which aims to in some way address this (e.g. high throughput screening, genomics, computer aided drug design) but the majority focusses solely on biological, chemical or physical factors. In reality, the success of drug development programs is not only governed by scientific factors, but also a wide variety of "external" factors, including intellectual property, the economic environment, funding, and many others. Incorporating such factors into analysis during drug development could help to improve decision-making.

Intellectual property, specifically patents, represents a particularly rich and accessible source of data[9]. Patents are an essential component of biomedical product development, are freely available, and are published in comparable numbers to academic research papers. In 2013, for example, 2.2 million patent applications[10] and 2.4 million academic articles[11] were published. Despite the remarkably similar numbers, in some cases there is little overlap: in an analysis of 2 million compounds extracted from journals and patents, just 6% of compounds in patents also appeared in journal articles[12]. Large scale attempts to mine this data for a variety of purposes exist (e.g. [13,14]), but within academia the prevalence of any analysis incorporating patents is very low[9].

Patent landscaping is an example of patent analysis that has, however, proliferated somewhat in academia. It is a methodology for analysing multiple patent documents for a variety of purposes such as uncovering technological trends[15] or examining opportunities for innovation[16]. The process for conducting a patent landscape is analogous to that of conducting a systematic review, but rather than academic literature, patent documents are identified and analysed. The process typically comprises four main stages: i) defining the purpose and scope, ii) designing and conducting searches, iii) data cleaning and curation, and iv) data analysis and interpretation[17] (Figure 1.1 illustrates this in more detail). The results of such analyses may be used for decision making in academia, industry and government, and, given the importance of patents in the drug development process, could be useful in efforts to reduce attrition.

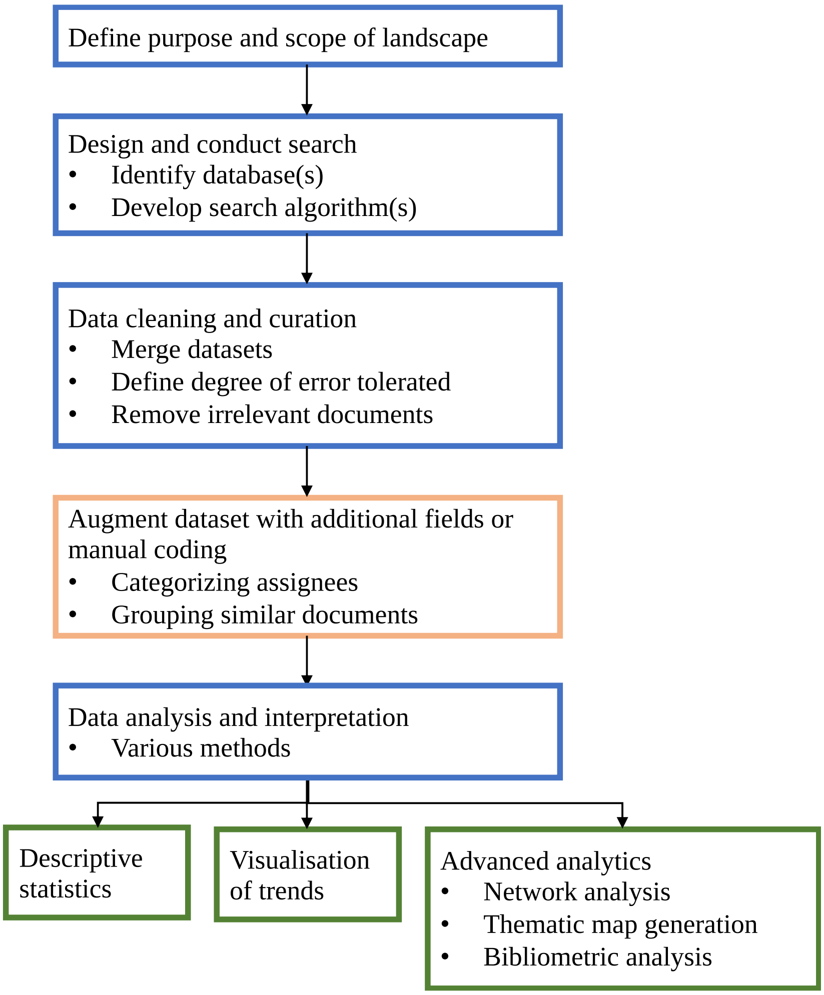

Define purpose and scope of landscape
Design and conduct search
Identify database(s)
Develop search algorithm(s)
Data cleaning and curation
Merge datasets
Define degree of error tolerated
Remove irrelevant documents
Augment dataset with additional fields or manual coding
Categorizing assignees
Grouping similar documents
Data analysis and interpretation
Various methods
Descriptive statistics
Visualisation of trends
Advanced analytics
Network analysis
Thematic map generation
Bibliometric analysis

Figure 1.1: Patent landscaping process

The typical process for conducting patent landscape analysis. In blue boxes are steps necessary for all patent landscapes, steps in the orange box are optional and the green boxes are example outputs. Adapted from[17].

When originally planning the research in this thesis, we therefore postulated that patent landscaping might represent an under-exploited approach to help to address the attrition problem. In reviewing the academic literature focussed on patent landscaping, however, it became apparent that improvements in the quality of available evidence were needed before such analyses could be useful. In particular, it was generally not possible to discern the exact methods used such that they could be replicated. To examine the extent and severity of this problem, we conducted a systematic review of the reporting quality of patent landscape articles, and, on the basis of the results, developed a guideline which aimed to improve it. The resulting work comprises Chapters 2 and 3.

Subsequently, we focussed attention more directly on the problem of reducing attrition during drug development. Despite dramatic increases in knowledge and throughput for techniques used in drug discovery, gains in output have not been realised, and many consider that an over-emphasis on reductionist approaches might be at least partly responsible[18,19]. We hypothesised that empirical approaches, incorporating external factors mentioned above in some way, might assist in identification of promising drug candidates. Regulatory approval is the “gold standard” reference point for drug development programs[18], so we aimed to develop approaches that could predict it. We systematically reviewed and assessed the quality and performance of existing approaches to do this in Chapter 4. Finding that only a narrow range of predictors had been explored and numerous methodological issues, we then developed models to predict approval in Chapter 5.

1.1 Structure of the Thesis

This thesis is structured as a set of largely self-contained chapters, each with independent introductions and discussions, followed by a general discussion chapter. Chapters 2 and 3 (with some modifications) are published or are accepted for publication. For Chapter 4, a protocol is published on PROSPERO and the work is complete. For Chapter 5, some minor additional work is suggested and potential future directions are provided. Chapter 6 is the general discussion chapter.

More specifically, the content of each chapter is as follows:

- Chapter 2 is a systematic review of reporting quality in academic patent landscape articles. The systematic review showed that reporting quality was insufficient and that methods were generally too poorly reported to allow reproducibility. The findings provided justification for the development of a reporting guideline. The majority of this chapter is published in Nature Biotechnology[20].

- Chapter 3 reports and describes the development of the Reporting Items for Patent Landscapes (RIPL) statement. The statement lists the essential items that should be reported in a patent landscape article. It is focussed on and developed for academic articles, though may also have utility for industry and government. It was developed *via* a consensus process involving input from patent landscaping and reporting quality experts. A paper resulting from this work has been accepted for publication in Nature Biotechnology and will appear in the November issue in 2018.

- Chapter 4 returns to the question of predicting regulatory approval by conducting a systematic review of existing attempts to do so. Only three relevant papers were identified. Methodological issues in two papers result in a very high likelihood of bias and a lack of reporting of the final model in the third means that the model cannot be used. A variety of predictor variables are included in the models, but none include physicochemical parameters of the drug candidates themselves. A systematic review protocol for this chapter is registered on PROSPERO[21].

- Chapter 5 presents the development and internal validation of two multivariable predictive models of regulatory approval for small molecules. The first includes a wide range of candidate predictor variables, including those related to literature (patents and academic), economics, the usage of the drug, and its physicochemical parameters. The second model includes fewer predictors, all of which are simple physicochemical descriptors, which allows a larger number of observations to be included in the dataset. The performance of both models is modest (area under the receiver operating characteristics curve, AUC~0.6), though the latter is marginally better. The model performs better than Lipinski's rule-of-5 in our dataset.

- Chapter 6 provides a general discussion and summary of key themes. We review reporting quality literature and some suggestions for future research in this area are made. The use of patents as an additional literature source in academia is then discussed. General criticisms that may be raised over our attempts to predict regulatory approval are then addressed and followed by a brief discussion of potential future work. We conclude with an overview of several themes that have emerged across the thesis.

2 Evidence of Insufficient Quality of Reporting in Patent Landscapes in the Life Sciences: A Systematic Review

Major components of this chapter first appeared in the following paper: "Smith JA, Arshad Z, Thomas H, Carr AJ, Brindley DA: Evidence of insufficient quality of reporting in patent landscapes in the life sciences. Nature Biotechnology, 210-214, 2017" by Springer Nature. Additional analysis of inaccessible articles and reporting quality over time is included, a new abstract and new figures are included, minor changes to content and typography have been made throughout and the document has been restructured.

2.1 Abstract

Patenting is an important part of the commercialisation process for life science and healthcare technologies. As a result, patent landscaping has emerged as a methodology to allow analysis and measurement of patent documents and patent activity, potentially leading to insights regarding research, innovation and development in technology areas. Despite the potentially important uses of such analyses, in this chapter we show that generally the quality of reporting in patent landscapes published in academic journals is low. A systematic review identified 81 patent landscaping papers which were evaluated against a checklist of items that should be reported in those papers. No papers complied with all of the checklist items, and very few (10%) complied with what we consider to be essential methodological items. Reporting quality improved

significantly over time but remains suboptimal. Impact factor was not correlated with reporting quality and research articles were marginally but significantly better reported than other article types. These findings identify a need for improved reporting in patent landscaping which we propose to address through the development of a reporting guideline.

2.2 Introduction

Patents in the life sciences are a critical metric of innovation and a cornerstone of the commercialisation of new life science and healthcare technologies. Patent landscaping has emerged as a methodology for analysing multiple patent documents in order to uncover technological trends[15], geographic distributions of patents[16], patenting trends and scope[22], highly cited patents[23], and a number of other uses[17]. Many such analyses are published in high impact journals[24–26], potentially gaining high visibility amongst academic, industry and government stakeholders. They may be used to inform decision-making, for example, prioritisation of funding areas, identification of commercial competition and therefore strategy development, or implementation of policy to encourage innovation or ensure responsible technology licensing. Patent landscaping may also provide a means for answering fundamental questions regarding the benefits and drawbacks of patenting in the life sciences, a subject on which there is considerable debate[27–29], but limited empirical evidence.

The aim of the patent landscaping process is to capture a set of pre-defined patent documents and to analyse them in some manner (definition: Box 2.1). The landscaping process (summarised in Figure 1.1) is analogous to the process for conducting systematic reviews of academic literature; however, in patent landscapes, the patent documents replace academic articles. Like systematic reviews, patent landscapes are important in allowing higher level insights to be drawn that could not be achieved by analysis of documents in isolation. Unlike systematic reviews, however, patent landscapes are often published under different guises than original research articles, possibly leading to a lack of emphasis on disclosing key information underpinning analyses and conclusions.

Box 2.1: Definition of a patent landscape

A landscape is an analysis of the relationships between multiple sets of indicators or of those indicators measured against temporal, technical or spatial dimensions. At least one of the indicators, in the case of a patent landscape, is patent publications or some aspect thereof. A landscape seeks to encompass an entire population of relevant data, rather than a random sample drawn from that population (adapted from[17]).

As in any research, to allow for reproducibility and effective evaluation of accuracy and quality, it is essential that studies are reported adequately. The specific methodologies used and results generated must be reported; clear justification and aims of investigations need to be provided in order to assess the validity of any conclusions; any funding or conflicts of interest should be disclosed to assess potential biases. The importance of adequate reporting has been heavily emphasised and policed in health research[30–32], including in systematic reviews[34]; however, outside of health research, discussion has been more limited. Despite the important potential implications of patent landscapes from a scientific, commercial and political perspective, criticism has emerged over a lack of standardisation, transparency and clear justification of methodology[35,36].

The purpose of this chapter is to provide the first systematic assessment of the quality of reporting in patent landscapes in the life sciences published in academic journals. Given the heterogeneous nature of methodologies used for, and questions addressed by, patent landscaping[35], consistent methodological practice may not be practical or useful, and could indeed stifle innovative methodologies from emerging. Clarity of reporting, on the other hand, is essential regardless of the approach taken and question investigated, and its assessment should represent a significant component of the peer review process and should be expected in scientific articles. Patent landscapes are published in a variety of journal types of various impact factors, and as various different article types (e.g. review articles and original research articles). The relationship between these indicators and reporting quality is also explored.

2.3 Methods

We follow the preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement recommendations[33,34] for methodology and reporting, where appropriate. No review protocol was registered.

2.3.1 Eligibility Criteria

Inclusion criteria for relevant articles were determined prior to any searches being conducted (summarised in Table 2.1). Articles were included if they stated somewhere in the article that a search for patent records was conducted, though were not included if they were purely a review of selected patent records not explicitly identified through a search. Some analysis of the included patents had to be conducted: for example, graphical or statistical; a list of patents could not simply be provided. The purpose of these criteria was to distinguish those studies that are landscapes (i.e. those including a search), and those which could be considered more similar to a narrative review (i.e. those not including a search). Only articles focussed on the life sciences were included, which was defined as "the sciences concerned with the study of living organisms, including biology, botany, zoology, microbiology, physiology, biochemistry, and related subjects"[37] with the addition of all medical sciences. Any article type or format which met the eligibility criteria and was published in a journal was considered relevant. This did not include book chapters or reports. No language restrictions or publication date restrictions were applied, though if an article could not be obtained in an English translation it was not included.

Table 2.1: Summary of eligibility criteria
All eligibility criteria had to be met for a study to be included

	Eligibility Criteria
1	The article included a statement that a search for patent records was conducted
2	The article included at least some analysis or interpretation of the identified patents, and was not simply a list of patent records
3	The subject matter of the article was the life sciences, which was taken to include the medical sciences

2.3.2 Search Strategy and Sources of Information

Search terms used to identify papers for screening are detailed in Supplementary Table 8.1. MEDLINE and EMBASE (via Ovid), PubMed, Scopus, and Science Direct were searched from their respective start dates until 16th March 2016. PubMed and Ovid were chosen because of their broad coverage, topically and geographically, of medical and life sciences research. Scopus and Science Direct were searched, in addition, to capture additional potentially relevant papers which scoping searches revealed did not appear in journals indexed by PubMed and Ovid; both databases are considerable in scope, including social science and humanities research outputs, in which it is possible that patent landscapes focussed on life sciences areas could be published.

2.3.3 Study Selection

Three stages of selection were applied: automatic duplicate removal, manual screening of abstracts and manual screening of full text papers. Duplicates were removed automatically using Mendeley Desktop (v1.16.3) software, and the resulting records were manually screened for relevance. At least two authors (JS, ZA and/or HT), independently screened abstracts of identified articles to determine relevancy. At the initial screening phase, articles for which there were discrepancies were included for full text evaluation. If discrepancies arose during full text screening, consensus was reached by discussion between the authors. Only articles which were available *via* one of the authors' institutions as full text articles were included for evaluation of reporting quality.

2.3.4 Reporting Quality Criteria and General Article Information

A data extraction sheet was developed for evaluation of reporting quality and collection of general article information. Criteria were selected for evaluating the quality of reporting based initially on the PRISMA checklist[34], which is a checklist developed for reporting of systematic reviews and meta-analyses. Some criteria were removed as they were not deemed relevant to patent landscapes. The authors then discussed and added additional criteria which might be required for transparency and reproducibility and updated the list. Apart from those criteria stated as being in “Title” and “Abstract” sections (Table 2.2) items were considered to have been reported regardless of the location in which they were reported, including supplementary material. This was because a number of papers (e.g. review articles) did not follow a structured format.

In addition, data were extracted on the type of analysis conducted, technology area under investigation, the article type, the SCImago Journal Rank (SJR) of the journal in which the paper was published, and general bibliographic information to allow summaries and trends to be determined (full list of items extracted in Table 2.2). The SJR was chosen rather than impact factor because historical data is readily available for SJR, and it is a ranking so corrections for time dependent trends in citations do not need to be made.

Table 2.2: Information extraction form

For all items other than the general article information, papers were coded in a binary system according to whether or not that item was reported (1 = reported, 0 = not reported).

Section	Item #	Item
General article information	A	Full bibliographic information
	B	Type of analysis conducted
	C	Technology area of investigation
	D	Article type
	E	SCImago Journal Rank (SJR) of the journal in which the article is published for the year in which it is published
Title	1	Article identified as a patent landscape

Section	Item #	Item
Abstract	2	Objective overview of aims, methods and findings provided
Introduction	3	Aims and rationale of the investigation are clearly stated
Methods	4	Clear description of the patent records aiming to be collected is provided
	5	Databases used to collect patent records are disclosed
	6	Date ranges for any searches conducted are provided
	7	Patent offices searched are specified
	8	Component of patents searched is stated (e.g. claims, abstract, title)
	9	Full electronic search strategy for at least one database searched is given
	10	Process for selecting relevant patents is outlined
	11	Software used for any analysis of data is detailed
	12	Details of any data analysis is provided
	13	Patent selection and/or data extraction, if applicable, is blindly reproduced
Results	14	Summary statistics for the dataset (e.g. in its simplest form, number of patents included in analysis)
	15	If data is extracted from individual patents, the data is included with the relevant patent citations
	16	Results of any statistical analysis conducted is included
	17	Patent records included in the study are listed, or a means to access them is provided (e.g. reference to supplementary material containing the list)
Discussion	18	The main findings of the study are discussed
	19	The limitations of the study are discussed
Conflicts of interest	20	Any conflicts of interest are stated, and sources of funding are disclosed

Table 2.3: Definitions of types of analysis used in extraction sheet

Analysis Type	Definition
Temporal	Any analysis that analyses patent records temporally

Analysis Type	**Definition**
Assignee	Identification or analysis of the assignees of at least some patent records in a dataset
Inventor	Identification or analysis of the inventors of at least some patent records in a dataset
Geographical	Any analysis that breaks down a dataset of patent records by geographical location
Citation	Analysis of forward and/ or backward patent record citations
Classification	Coding of patent documents according to the technical features of their content, as per generally recognised coding systems such as the Cooperative Patent Classification (CPC) system
Cluster mapping	The use of visual display to highlight recurring themes and key words within patent documents and their relationships
Patentability	An analysis into the patentability of certain subject matter
Freedom to operate	An analysis that aims to identify potential freedom to operate issues, which was taken to include any assessment of potential “blocks” to innovation, invention, etc
Patent watch	A summary or analysis of newly issued patent records, usually applications
Validity	An analysis into the validity of patent claim(s)
Active/ inactive	An analysis of the patent status (active or inactive) for one or more patent records
Application/ granted	An analysis of application status (application or granted) of patent records
Claims analysis	Any detailed analysis of claims, when stated in the context of being a claims analysis

2.3.5 Data Extraction

At least two authors (JS, ZA and/or HT), independently extracted data from full text articles using the information extraction form (Table 2.2). For all items apart those in the section general article information, items were assigned a binary outcome variable of compliance (1) or non-compliance (0) for each article. To ensure objectivity, an article was considered to be compliant if any attempt whatsoever was made to report relevant information; for example, in the discussion section, if any limitation was mentioned, the article was considered compliant with that criterion. Any discrepancies in scoring were discussed and resolved amongst the authors. To assess SJR, a search for the relevant journal was carried out on the SJR website (www.scimagojr.com) and SJR was recorded for the year of article publication. Definitions used to determine analysis type are provided in Table 2.3. Article type was determined based on the category in which the journal placed the article. If this was unavailable, articles were allocated to a category based on their internal identification, e.g. if it was stated that the article was a review of the literature, it was considered a review. If this was in turn unavailable, the authors discussed the category into which they believed it most closely fitted and allocated it once consensus was achieved. The Thomson Reuters' life science definitions were used for categorisation of articles into different areas (http://mjl.clarivate.com/scope/scope_ccls/) though the definition of "Animal and Plant Sciences" was expanded to include husbandry. Any discrepancies in opinion were discussed between two authors until consensus was reached.

2.3.6 Statistical Analysis

Some items included in the extraction form were not relevant to every paper (e.g. item 13), so for each paper a percentage compliance was calculated based on the number of compliant items and total number of relevant items to allow comparability between papers. Statistical tests were conducted in R (v3.3.1) and graphs were generated in R or Microsoft Excel (v16.17). Comparisons were made between mean article compliance by article and journal type. Spearman's rank correlation coefficient and associated significance test was calculated in R to test for correlation between impact factor and percentage compliance and impact factor and year of publication. An alpha of 0.05 was used and a Bonferroni correction applied to account for multiple comparisons. Therefore, results were considered significant when $p < 0.0125$.

2.4 Results

2.4.1 Search Results, Study Inclusion and Study Characteristics

Searches and screening resulted in a total of 81 full text studies for inclusion in this systematic review (Figure 2.1; citations in Supplementary Table 8.2). After removal of duplicate records, the abstracts of 3,348 articles were screened for relevance, and 577 records remained which could not be excluded based purely on review of abstracts and titles and which were assessed as full texts, where available. 380 articles were excluded because upon deeper examination they were not patent landscapes or reviews (n = 201), there was no mention of a search for patents (n = 78), or they were not life science focussed (n = 100). A further potentially relevant 116 articles were identified which we were unable to exclude based purely on abstract or title, but for which the full text was not available at the time of analysis.

Full text papers included for analysis conducted a broad range of analyses (Figure 2.2A) and examined a number of different research areas (Figure 2.2C). The most common forms of analyses were temporal, assignee, and geographical, with each type of analysis appearing in 75% or more papers. Nearly half of the papers included in the study were research papers (49%), followed by reviews (29%), and other article types (22%; Figure 2.2B). The majority (73%) of papers were published in scientific journals, with the rest falling into other categories (23%) such as legal or business journals (Figure 2.2D).

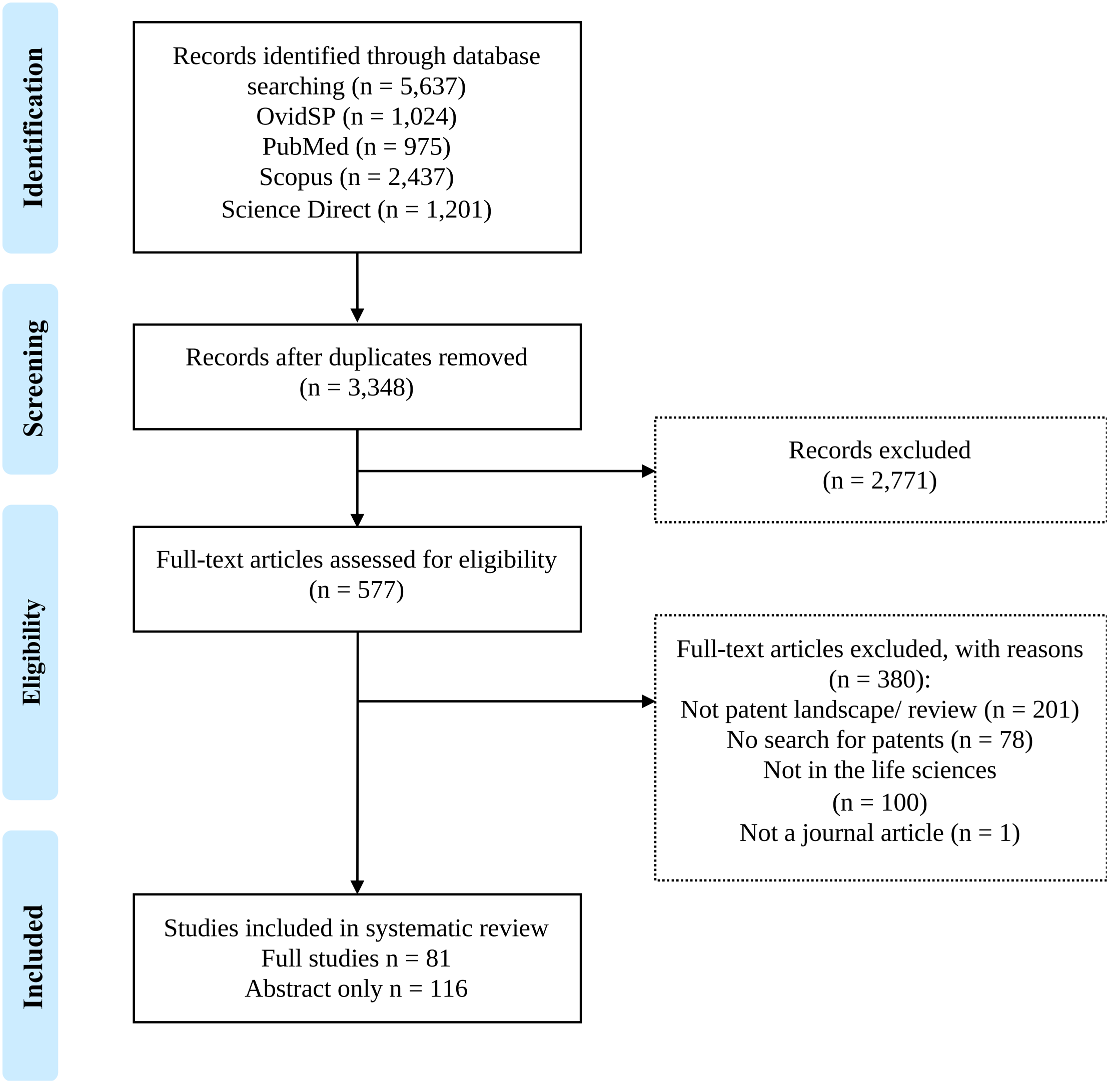

Identification
Records identified through database searching (n = 5,637)
OvidSP (n = 1,024)
PubMed (n = 975)
Scopus (n = 2,437)
Science Direct (n = 1,201)
Screening
Records after duplicates removed (n = 3,348)
Records excluded (n = 2,771)
Eligibility
Full-text articles assessed for eligibility (n = 577)
Full-text articles excluded, with reasons (n = 380):
Not patent landscape/ review (n = 201)
No search for patents (n = 78)
Not in the life sciences (n = 100)
Not a journal article (n = 1)
Included
Studies included in systematic review
Full studies n = 81
Abstract only n = 116

Figure 2.1: PRISMA flow diagram

PRISMA flow diagram[34] *detailing number of studies included at each stage and reasons for removal.*

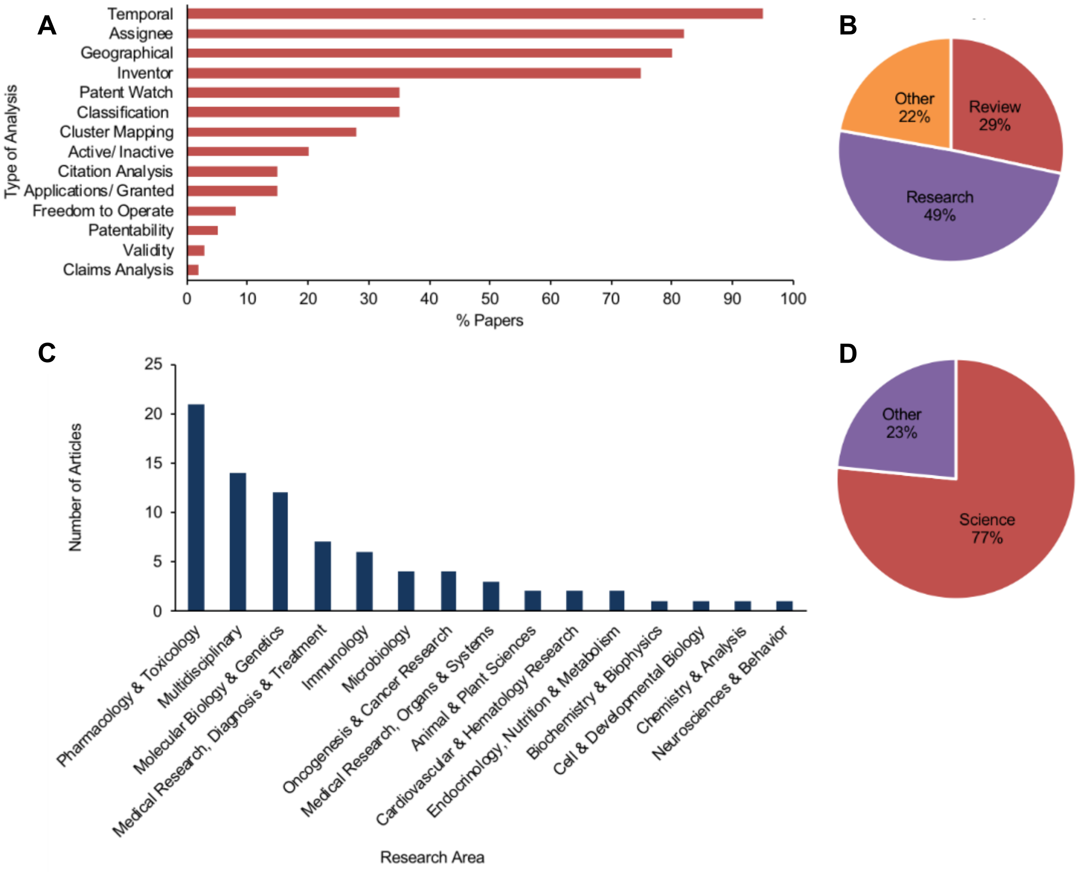


Figure 2.2: Characteristics of included papers

A) Types of analysis: *percentage of patent landscaping articles conducting different types of analysis in studies included in this systematic review (n = 81 for all). Each article could be assigned more than one type of analysis (definitions: Table 2.3).* ***B) Article type:*** *article types for included papers.* ***C) Technology area:*** *technology areas focussed on in the articles included in this systematic review. Each article was assigned to a single technology area.* ***D) Journal type:*** *type of journal in which patent landscapes were published: scientific or other, which included journals primarily focussed on legal, business, or other topics.*

2.4.2 Quality of Reporting

In 81 articles assessed for compliance against 20 items considered together to represent an adequately reported study, no articles reported all of the items included in our checklist that were relevant to them. Mean compliance across all articles was 64% (SD±15%). Table 2.4 lists the number and percentage of articles reporting each item.

Percentage compliance was not normally distributed (Shapiro-Wilk test of normality, $W = 0.93392$, $p < 0.001$) and normality could not be achieved through transformation of the data. Therefore, Kruskal-Wallis tests were used for comparisons of compliance against article and journal type, and Spearman's rank correlation coefficient was calculated for compliance against SJR and year. No correlation between SJR and compliance was seen (Spearman's Rho, $r_s = 0.085$, $p = 0.475$; Figure 2.3A). Mean compliance in research, review and "other" articles were 70% (SD±8), 55% (SD±16) and 60% (SD±17), respectively (Figure 2.3B). A significant effect of article type on compliance was observed (Kruskal-Wallis Test, $H = 20.5$, $p < 0.001$). Mean compliance in scientific and "other" journals was 64% (SD±16) and 63% (SD±11; Figure 2.3C), respectively, which did not differ significantly (Kruskal-Wallis Test, $H = 0.114$, $p = 0.736$). A significant positive relationship between year of publication and compliance was seen (Spearman's Rho, $r_s = 0.409$, $p < 0.001$; Figure 2.3D).

Table 2.4: Number and percentage of articles reporting items in the reporting quality checklist (n = 81)

Section	Item #	Item	Articles Reporting (%)
General article information	A	Full bibliographic information	N/A
	B	Type of analysis conducted	N/A
	C	Technology area of investigation	N/A
	D	Article type	N/A
	E	SCImago Journal Rank of the journal in which the article is published for the year in which it is published	N/A
Title	1	Article identified as a patent landscape	46 (56.1)
Abstract	2	Overview of aims, methods and findings provided	36 (43.9)
Introduction	3	Aims and rationale of the investigation are stated	81 (100)
Methods	4	Description of the patent records aiming to be collected is provided	78 (95.1)
	5	Databases used to collect patent records are disclosed	75 (91.5)
	6	Date ranges for any searches conducted are provided	61 (74.4)
	7	Patent offices searched are specified	65 (79.3)
	8	Component of patents searched is stated (e.g. claims, abstract, title)	43 (52.4)
	9	Full electronic search strategy for at least one database searched is given	35 (42.7)
	10	Process for selecting relevant patents is outlined, if applicable	39 (55.7)*
	11	Software used for any analysis of data is detailed	50 (61.0)
	12	Details of any data analysis are provided	73 (89.0)
	13	Patent selection and/or data extraction, if applicable, is blindly reproduced	1 (1.4)*
Results	14	Summary statistics for the dataset (e.g. in its simplest form, number of patents included in analysis)	66 (80.5)

Section	Item #	Item	Articles Reporting (%)
	15	If data is extracted from individual patents, the data is included with the relevant patent citations	45 (54.9)
	16	Results of any statistical analysis conducted are included	68 (82.9)
	17	Patent records included in the study are listed, or a means to access them is provided (e.g. reference to supplementary material containing the list)	20 (24.4)
Discussion	18	The main findings of the study are discussed	78 (95.1)
	19	The limitations of the study are discussed	38 (46.3)
Conflicts of interest	20	Any conflicts of interest are stated and sources of funding are disclosed	37 (45.1)

**n = 70*

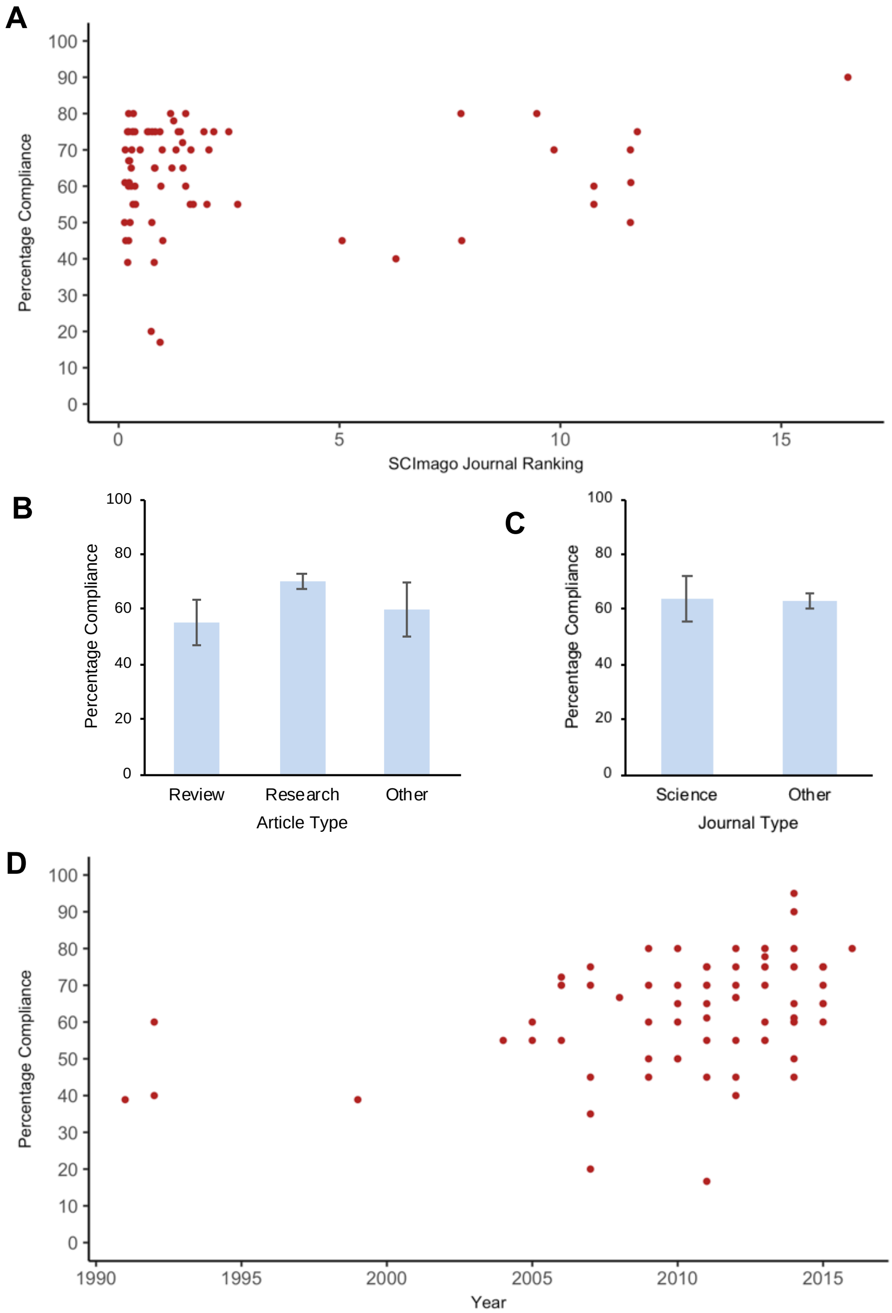


Figure 2.3: Reporting quality by journal ranking, article type, journal type and year of publication (n = 81)

Compliance with reporting checklist vs. A) SCImago Journal Ranking: *no relationship between SJR and percentage compliance to the reporting item checklist (Spearman's Rho,* r_s =

*0.085, $p = 0.475$); **B) Article type:** significant effect of article type on compliance (error bars ± 95% CI; $n = 81$; $H = 20.5$, $p < 0.001$); **C) Journal type:** no relationship between article type, (scientific journals and "other" journal types, such as legal or business journals) and percentage compliance (error bars ± 95% CI; Kruskal-Wallis Test, $H = 0.114$, $p = 0.736$). **D) Year:** significant positive correlation between year and percentage compliance (Spearman's Rho, $r_s = 0.409$, $p < 0.001$).*

2.4.3 Inaccessible Articles

We were not able to include 116 potentially relevant articles due to lack of availability of the full text articles. For these articles, we collected the year of publication and journal type. Year of publication was similar in the inaccessible articles (median = 2011, IQR [2008, 2013]) to those in the included sample (median = 2012, IQR [2009, 2014). The majority (85%) of inaccessible articles, however, were from journals of type "other", whereas in the included sample the majority (77%) were "science" (Figure 2.2D). Our analysis of included articles showed a significant relationship between year of publication and reporting quality (Figure 2.3D) and no effect of journal type on reporting quality (Figure 2.3C), so we do not have reason to believe that the inaccessible articles would differ considerably in their reporting quality to accessible articles.

2.5 Discussion

The final step in conducting academic research is not the analysis of data, but the communication and dissemination of the research in a manner that allows the approach taken to be determined and reproduced, and the validity of the findings to be assessed[38]. The primary mechanism of this dissemination and communication within academia is *via* journal publications, which represent the means by which the quality of a study is assessed. Our analysis provides evidence that the quality of reporting in patent landscapes in the life sciences in such publications is insufficient.

2.5.1 Reporting of Methods and Results

Many articles do not report critical methodological items. Without full reporting of the following methodological items: eligibility criteria of patents to be included; search strategy; databases, dates, patent offices and components of patents searched; software used; patent selection process; and details of analysis conducted (items 9 to 17, Table 2.4), reproducing and validating a patent landscape methodology is, in our opinion, not possible. Eight articles (9.9%) reported all of these methodological items that were applicable to them. Just one article (1.4%) reported that patent selection was blindly reproduced, and 24% of articles list the patents included in the study: the dataset upon which conclusions and analyses are based.

The lack of reporting of key methodological and results items challenges interpretation of conclusions and renders reproduction or updating of studies impossible in many cases. A patent landscape represents a considerable investment of effort and time, and without clarity over its methods and results, the impact of this effort can be significantly reduced. Greater clarity would allow additional analysis to be conducted by future researchers, maximising the potential benefit of the research[39]. Improvements in methods and results reporting might allow for meta-analyses of patent landscaping papers and allow independent researchers to use results in their own investigations for other purposes.

2.5.2 Potential for Conflict of Interests

Conflicts of interest and funding sources is also a relatively poorly reported item (45%). Patents are inherently linked to commercial interests and, therefore, the declaration of conflicts of interest and any funding sources is important in the evaluation of potential biases. In clinical trials, lower reporting quality has been associated with increased effect sizes[40]. While there is no effect size, as such, to evaluate in patent landscaping articles, incentives for the portrayal of inaccurate information by authors or funders could easily be conceived, and improved transparency over conflicts of interest and funding should therefore be encouraged. This issue is compounded by the fact that reporting of methods and results often challenges the ability to determine accuracy of information.

2.5.3 Explaining Reporting Quality

Our findings are not limited to lower quality journals. It is generally assumed that the higher the quality of, and therefore the quality of reporting in, an article, the higher is the likelihood that it will published in a high-impact journal[38,41]. However, no statistically significant correlation was observed between SCImago Journal Rank and the quality of reporting in patent landscaping articles (Figure 2.3A) suggesting that reporting quality is not currently an important determinant in the publication of patent landscapes in high quality journals. This is unlike other areas, in which a number of studies conducting similar analyses have found significant relationships between impact factor and reporting quality[42,43], and may be due to the fact that there are a limited number of well reported patent landscapes on which to model reporting.

Small but significant differences in reporting quality are associated with different article types. Perhaps unsurprisingly, reviews are less well reported than research articles (Figure 2.3B), which indicates that structured paper formats might be more conducive to complete reporting. Given that patent landscapes may just be one component of an otherwise narrative review, structured reporting, as in many research articles, is unlikely to be practical for all papers. However, emphasising that the patent landscaping component of any paper represents research that others may wish to rely on for further analysis could help to improve reporting, and details could be reported in supplementary materials.

A significant increase in reporting quality with year of publication was seen (Figure 2.3D), a promising finding consistent with research in other areas[44–46]. This improvement may reflect a general trend of greater emphasis on transparency and disclosure[47,48] or the efforts of other analyses and critiques of patent landscape articles[17,36]. Though promising, recent papers still do not report many essential items and reporting quality overall remains suboptimal.

2.5.4 Analysis Types

Patent landscaping has previously been described, in the most part, as relatively simplistic[17], and much analysis recorded and observed herein supports this. A large proportion of papers present numbers of patents over time, patents per geographical region, patents per assignee or inventor, or other count data (Figure 2.2C). There is no inherent problem in the use of such analyses; however, patent numbers and other such measures cannot alone portray all of the information that may be useful or required by the reader for appropriate interpretation[49]. For example, it is possible that within one research area, there are a large number of patents focussing on a very narrow set of inventions; another research area may have very few patents of broad scope. Without some consideration of the information within the patent documents, discerning these differences would be challenging. With appropriate discussion of limitations, these issues could be allayed; however, limitations were discussed in less than half of the included papers (46%).

More advanced analytical approaches are employed in some cases to interrogate patent documents in more detail, though these still raise some concerns. Software is used to thematically cluster patent documents and present the outcome visually in “cluster maps” (28%). From such data, areas of high patent activity are often identified (e.g. [15,22,50]), and may be used to identify “gaps” in the technology or research landscape[51]. Other papers similarly mention the use of patents to identify gaps, though not through an explicit methodology[52,53]. Proprietary software is often used to generate cluster maps, but the algorithms underpinning such software are rarely discussed or detailed in the papers using them, and in some cases do not appear to be publicly available at all. Without detailed analysis of patent claims by an expert, the identification of gaps in research or technologies is difficult to ascertain with existing methods. However, very few papers conducted any form of claims analysis (2%). As above, appropriate discussion of limitations would go some way to addressing these concerns.

2.5.5 Outlook

The findings of this systematic review are congruent with similar investigations that have been conducted in other fields which report omissions in methods[54,55], including statistical methods[56], incomplete presentation of data preventing later analysis[57], and inadequate conflict of interest reporting[58]. The findings also provide empirical evidence for statements previously made in relation to the quality of patent landscapes[17]. To address reporting issues in other fields, a great number of reporting guidelines have been developed which provide checklists of items that should generally be reported and which together represent an adequately reported study. Introduction and endorsement[59] of guidelines has been associated with improved reporting quality of clinical trials. This chapter provides the empirical justification for the development of a guideline to improve the quality of reporting in patent landscaping articles, which is registered on the Equator Network website (www.equator-network.org), and which is reported in the following chapter. The purpose of the guideline is to improve transparency and standardisation of reporting, in order to allow reproducibility, comparability, and accurate evaluation of patent landscapes.

2.5.6 Limitations

This systematic review focussed solely on the reporting quality of patent landscapes in academic journals. It should be noted that patent landscapes are commonly reported outside of academic journals, such as in those conducted by government bodies or industry. The quality of reporting in available government publications (e.g. [60]), in general, appears to be quite good, perhaps due to the lack of limitation with regards to document length in comparison to academic papers. However, even seemingly detailed landscapes lack full disclosure of search terms[61,62] and contain only superficial explanations of the algorithms employed[63]. The findings of this study may still, therefore, be useful in the context of reporting studies outside of academia. We do not have access to proprietary patent landscapes conducted in industry, so cannot comment on the relevance of these findings to that audience.

Additionally, we were unable to access the full texts of 116 potentially relevant articles which could not be excluded based on abstract alone. Year of publication was similar for inaccessible and included articles, though the proportion of different journal types was different. Reporting quality in included articles varied with year of publication but not journal type, so we do not think that inclusion of these articles would have significantly changed the reported results.

2.5.7 Conclusion

Patent documents are an exceptionally rich source of information which can and should be mined and analysed for a number of purposes. The breadth of possibilities for analysis of patent documents may preclude the development of standardised methodologies, and as such this may not be possible. However, without adequate reporting, the full value of such analyses will not be realised, and even the most rigorous and elegant investigations may be limited in reach because they simply cannot be reproduced and critically evaluated.

3 Reporting Items for Patent Landscapes: The RIPL Statement

Major components of this chapter have been accepted for publication in Nature Biotechnology, by Springer Nature, and will appear in the November issue in 2018. The author list for the accepted publication is: Smith JA, Arshad A, Trippe A, Collins GS, Brindley DA, Carr AJ. A major additional section (Section 3.5.2, Explanation of Reporting Items) has been added to this chapter. It has also been restructured in comparison to the paper, several figures have been reformatted, a new abstract has been written and minor typographical changes have been made throughout.

3.1 Abstract

Reporting quality in academic patent landscapes is not sufficient to allow reproduction, validation or synthesis of existing work. The findings of patent landscapes could have implications for academia, industry, and government, so it is important to address this deficiency. We therefore developed a reporting guideline: the Reporting Items for Patent Landscapes (RIPL) statement. A two-round modified Delphi study was conducted *via* web survey to reach consensus on a minimal list of items to be reported. Participants had academic or commercial experience with patent landscape articles or had previously been involved in the development of guidelines for reporting of research. Following the study, reported items were consolidated by the authors into a 21 item checklist, considered essential to report in any academic patent landscape article. Explanation of these items is provided. Though the checklist is aimed at academic articles, it may also be useful for those produced by government or industry.

3.2 Introduction

Patent landscapes collate and analyse information from patent publications for a variety of purposes, such as analysing technological progress[15], identifying innovation gaps[51], or studying patenting practices[64]. They are a resource increasingly available to researchers and decision-makers in the life sciences (Figure 3.1). Many patent landscapes gain widespread attention through publication in high impact journals[24–26,65,66], potentially viewed by key stakeholders in academia, government and industry. Despite their potential importance, we have shown that the quality of reporting in patent landscape articles is generally insufficient[20], potentially leading to an inability to critically appraise, interpret, and synthesise findings. Other studies have called for greater harmonisation in standards of disclosure[35] and establishment of consistent practices and reporting criteria[17].

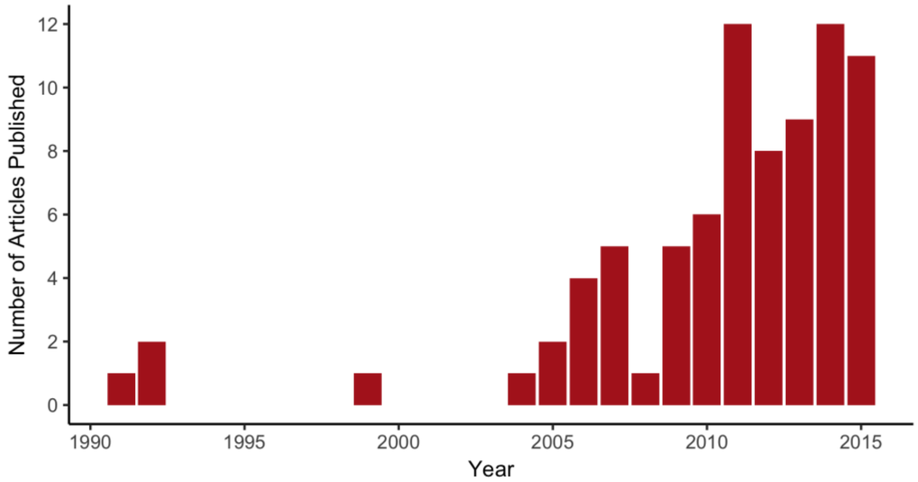


Figure 3.1: Number of patent landscape articles in academic journals increases

Number of patent landscapes published in academic journals per year is increasing steadily. Data are from the full text articles included in our systematic review in Chapter 2 (n = 80, 2016 paper removed due to incomplete data for that year).

Inadequate reporting has been widely reported in publications of health research[54,67,68], including economic assessments[69] and preclinical research[70], and a large number of reporting guidelines have been developed to improve reporting quality therein (e.g. [30,34,71,72]). Reporting guidelines have also been developed for specific methodologies, including microarrays[73] and data analysis in metabolomics[74], though such guidelines are much less prevalent. These and other guidelines can be found in the EQUATOR Network library of reporting guidelines (www.equator-network.org), where the development of this guideline was also registered.

Evidence suggests that the introduction of reporting guidelines can improve reporting quality in health research. For example, a systematic review of the impact of the Consolidated Standards for Reporting Trials (CONSORT) statement concluded that journal adoption of the statement is associated with improved reporting of randomised controlled trials[75]. A Cochrane review of the same subject similarly concluded that journal endorsement of the CONSORT statement may improve completeness of reporting[59].

To address the deficiencies in reporting quality of patent landscapes, we have developed a reporting guideline detailing minimal information that should be included in patent landscaping articles: the Reporting Items for Patent Landscapes (RIPL) statement. The guideline was developed using a modified Delphi protocol with input from experts in patent landscaping internationally.

3.3 Methods

A modified Delphi study was conducted to achieve consensus on a checklist of items to be included in a reporting guideline for patent landscapes. We did not, *a priori*, limit the number of rounds of the Delphi study; rather, it continued until one of the stopping criteria was met (detailed below). The modified Delphi study was followed by a phase in which the items selected for inclusion were consolidated into a final checklist.

3.3.1 The Delphi Process

The Delphi process can be described as an "exercise in group communication that brings together and synthesises the knowledge of a group of geographically scattered participants who never meet"[76]. The aim is to achieve consensus on a topic or series of questions through structured communication between a panel of experts[77]. Successive rounds of questioning are used, with information gathered from each round communicated back to participants in a standardised format. The process ends when consensus is achieved, or opinion fails to change between rounds.

3.3.2 Sample Identification and Selection

During the systematic review of patent landscapes described in Chapter 2, email addresses of first and senior authors of included manuscripts were noted where available, creating an initial list of experts suitable for inclusion. The authors also suggested potential experts for inclusion in the study and supplementary internet searches were conducted for individuals with relevant expertise. Individuals were contacted if they met the inclusion criteria (Table 3.1).

Table 3.1: List of inclusion criteria applied to select experts for inclusion in Delphi study

Participants had to meet at least one criterion

	Inclusion Criteria
1	Author on a published patent landscape study
2	Involved in production of patent landscapes in a commercial setting
3	Previous experience in development of guidelines for the reporting of research

Consensus on the number of individuals that should be recruited to a Delphi study is lacking[78]; however, the minimum number is generally around ten[79], and the median number of participants involved in developing reporting guidelines is approximately 22 (ref: [80]). Given that we expected the number of experts available to be lower than in some other areas of guideline development, and in anticipation of response rates between Delphi rounds of approximately 90% (ref: [76]), we aimed to recruit 20 participants. Once 20 participants had been recruited, the Delphi study was commenced.

Identified experts were invited to participate through a standardised email template, which included a study information sheet and participant consent form. Individuals were sent reminder emails one and two weeks after initial invitation; if they failed to respond, there was no further communication.

3.3.3 Questionnaire Production and Structure

The initial survey round represented an over inclusive list of reporting items that could theoretically be reported in a patent landscaping study (Supplementary Table 8.3). These were agreed upon through searches of the extant literature, including other reporting guidelines (in particular the PRISMA statement[34]), results of our previous systematic review, and consultation with experts. Questionnaires were pre-tested by an individual not involved in the initial production of the questionnaire prior to final distribution to ensure face validity and flow[81,82].

Each questionnaire consisted of seven themed sections based on the format of a standard research paper to facilitate the placement of items into appropriate categories in the final reporting guideline. These themes were: title, summary/abstract, introduction, methods, results, discussion and other. Each question consisted of a potential item for inclusion in the guidelines. Respondents were asked to rate the item on a ten-point Likert scale; ten and one denoted the participant strongly agreed or strongly disagreed with the item's inclusion, respectively. An answer of 1-4 or 7-10 indicated that the item should be excluded or included in the guidelines, respectively. An answer of 5 or 6 denoted that the respondent was unsure.

To gather qualitative data, space for free text responses was provided at the end of each section. Suggestions for new items to be added to the survey, and reasoning for specific responses, could be provided here. Participant occupation and location were also recorded.

3.3.4 Data Collection

The purpose of each round of the survey was to reach consensus on proposed items and to gather qualitative responses to determine if any additional items should be considered for the following round.

Once an individual agreed to participate, an email, including a link to complete the survey, was sent in which they were asked to complete the survey within two weeks. If they failed to complete the survey in the allocated time they were sent two reminders. Failure to respond within a fortnight of the initial reminder resulted in exclusion. There was a minimum five-week period planned between each survey round to allow time for data analysis, survey construction, and pre-testing.

When consensus was reached on an item, it was removed from the subsequent round of the questionnaire. In the subsequent round, items on which consensus was achieved were fed-back to respondents as were descriptive statistics for items on which stability and/or consensus had not been reached (see Section 3.3.5). This process was repeated until stability and/or consensus was reached on all items, until response rates fell below the critical value of ten individuals, or until the number of items on which consensus was not reached became impractically low (less than five). Questionnaires were delivered and responses collected in SurveyMonkey (www.surveymonkey.com).

3.3.5 Data Analysis

We considered consensus to be achieved when item responses had an inter quartile range of two on a ten-point Likert scale in one round of the survey[83,84]. For an item to be included in or excluded from the checklist, greater than 75% of experts had to agree that it should be included (a value of 1-4) or excluded (a value of 7-10). When question items met these criteria, they were not included in the next Delphi round.

Stability between rounds was assessed using the kappa test. Moderate consensus (0.41 or above) represented stability, based on Cohen's suggested interpretation[85]. If responses were stable between two rounds and consensus had not been reached, the authors discussed the items and determined whether or not they should be included.

Descriptive statistics (the mean, mode, median, range, standard deviation, and interquartile ranges) were calculated for each item in each round. The mean, standard deviation, median, and range were reported back to respondents. These parameters were chosen as they are universally understandable and provide a means for respondents to quickly assess the results of survey rounds. Free text responses were manually reviewed and if additional reporting items were suggested they were included in the following round. All statistical analyses were carried out on R (v3.2.1).

3.3.6 Round One

The initial survey round represented an over inclusive list of 57 reporting items that could theoretically be reported in a patent landscaping study (Supplementary Table 8.3). These were agreed upon through searches of the extant literature, including the PRISMA statement[34], results of our previous systematic review[20], and discussion amongst the authors.

3.3.7 Subsequent Rounds

Subsequent rounds presented to participants any items on which consensus had not been reached, and any new items which had been suggested in the previous round.

3.3.8 Consolidation of Consensus Items

Upon completion of the Delphi study, we consolidated the reporting items on which consensus had been achieved into a checklist. Because having a large number of reporting items in a checklist is likely to reduce its practicality and usability, and because reporting guidelines include a median of 21 checklist items[86], we aimed to generate an approximately 20 item checklist, grouping related reporting items into single checklist items. The final checklist was checked for comprehension and unambiguity by co-authors and revised where required.

3.3.9 Ethical Approval

The Central University Research Ethics Committee (CUREC) granted ethical approval to carry out the modified Delphi study under reference R46326/RE001.

3.4 Results

The modified Delphi study was conducted in late 2016 and early 2017, resulting in 48 items to be included in a consolidated reporting guideline for patent landscapes (Figure 3.2). The criteria for ending the Delphi study were met after two rounds.

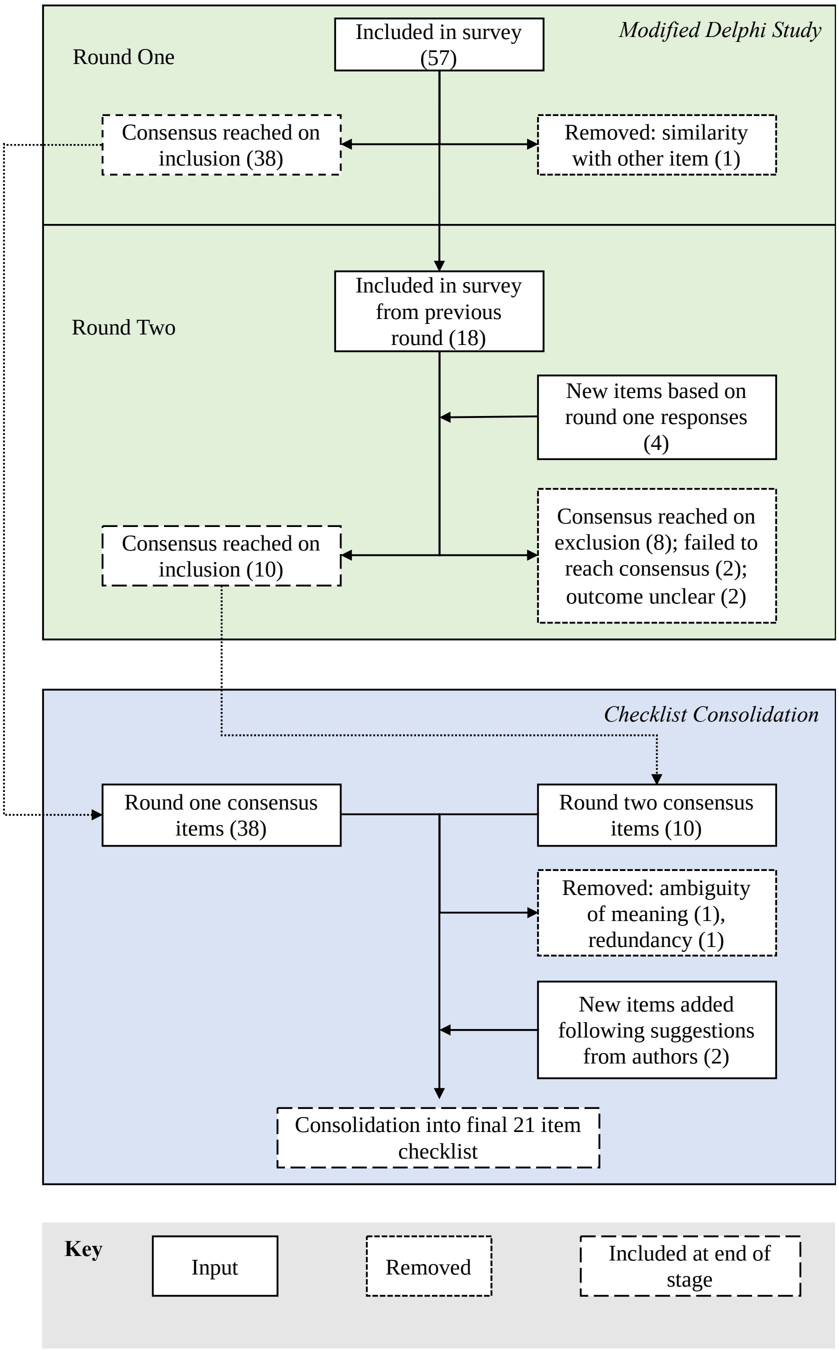

Modified Delphi Study
Round One
Included in survey (57)
Consensus reached on inclusion (38)
Removed: similarity with other item (1)
Round Two
Included in survey from previous round (18)
New items based on round one responses (4)
Consensus reached on inclusion (10)
Consensus reached on exclusion (8); failed to reach consensus (2); outcome unclear (2)
Checklist Consolidation
Round one consensus items (38)
Round two consensus items (10)
Removed: ambiguity of meaning (1), redundancy (1)
New items added following suggestions from authors (2)
Consolidation into final 21 item checklist
Key
Input
Removed
Included at end of stage

Figure 3.2: RIPL checklist development flow diagram

The checklist development was split into two main stages: a two-round modified Delphi study which aimed to achieve consensus on a number of proposed checklist items; and a subsequent stage where those items on which consensus was achieved were consolidated into a practical checklist. Number of items is given in brackets.

3.4.1 Participants

Twenty participants agreed to participate in the first round of the Delphi study, representing a participation rate of 29%. Of the remaining experts who were contacted, 11 declined to participate and 39 failed to respond. One respondent agreed to participate in the study but did not complete the first round, and another completed the first round but did not complete the second round (Figure 3.3). Participants were of a range of geographical locations, backgrounds and ages (Supplementary Figure 8.1) and both male and female.

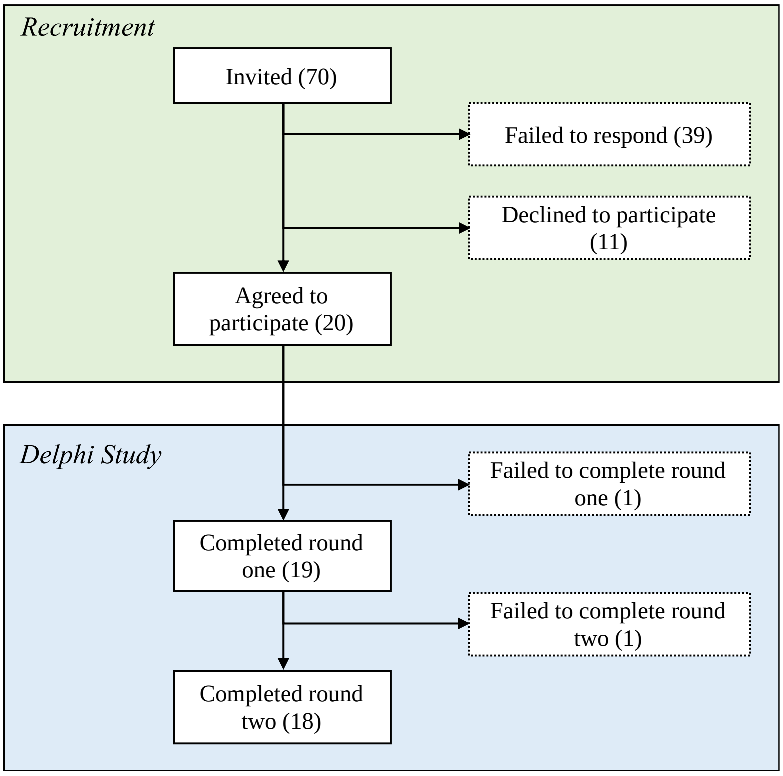

Recruitment
Invited (70)
Failed to respond (39)
Declined to participate (11)
Agreed to participate (20)
Delphi Study
Failed to complete round one (1)
Completed round one (19)
Failed to complete round two (1)
Completed round two (18)

Figure 3.3: Participant recruitment and participation

Number of experts is given in brackets.

3.4.2 Round One

In round one, consensus was reached on the inclusion of 38 items so these items were not included in the subsequent round, and one item was removed due to similarity in meaning to another item. No items were excluded. Consensus was not reached on 18 candidate items and these were included in the second round.

3.4.3 Round Two

During round one, it was noted by some participants that, although we listed each item within a certain section (e.g. abstract, introduction, etc.), in some cases it may be more appropriate to include information in supplementary material (e.g. item no. 15, list of patent numbers). Some participants also noted that patent landscapes are not always written as structured research papers. We therefore clarified at the beginning of round two that the sections are provided for the purpose of providing structure to the reporting guideline, and that in practice most items could be included in any part of the paper, or alternatively in the supplementary material.

The second round included the 18 candidate items on which consensus was not reached in the first round, in addition to four new candidate items added as a result of qualitative responses provided by survey participants during round one, leading to a total of 22 candidate items. The group mean, standard deviation, median, and range of scores for each item also present in round one were reported back to respondents. Consensus was reached on the inclusion of ten items and the exclusion of eight items. For two items, consensus was reached but the inclusion or exclusion criteria were not met, and thus the outcome was unclear. For two further items, consensus could not be achieved but responses were stable between rounds. As there were only two items on which consensus was not reached, no further Delphi rounds were completed.

In addition, based on comments received during the first round of the Delphi study, we proposed a definition for a patent landscape in the second round which was assessed for agreement and consensus in the same manner as reporting items, and on which consensus and agreement was reached.

3.4.4 Consolidation of Consensus Items

At the end of round two, consensus had been reached on the inclusion of 48 items in the reporting guideline. We consolidated the consensus items into a 19 item checklist, grouping related items into single items. 46 of the 48 items on which consensus was reached are represented in the final checklist; two items were removed due to ambiguity of meaning. The two items for which the outcome was unclear and the two for which consensus was not reached were excluded. Two additional items were added in the consolidation phase following suggestion by co-authors, resulting in a final 21 item checklist. The final checklist was checked for comprehension and unambiguity by all authors and revised where required.

3.4.5 Final Checklist

The final checklist of recommended reporting items is presented in Table 3.2. Explanation of each reporting item is provided in Section 3.5.2.

Table 3.2: Checklist of items to be reported in a patent landscaping article

Item # ***Section***	**Topic**	**Checklist Item**
Title		
1	Title	Identify that the article includes a patent landscape and state the subject matter under investigation (e.g. gene editing technologies)
Summary/Abstract		
2	Abstract	Provide a summary which includes the background, rationale, results and main findings in the context of the aims
Introduction		
3	Rationale	Describe the rationale for the study, including relevant background information and the potential impact of the investigation
4	Aims	Describe the aims of the study
Methods		
5	Search	State the databases and patent offices searched, the dates on which the searches were conducted, and the components of the patents searched. Include the search terms used for all databases searched
6	Selection criteria	Include details of the selection criteria of patents to be included in the patent landscape, including the subject matter of those patents
7	Identification of relevant patents	If applicable, state how patents identified in searches were sorted for relevance
8	Data extraction	List and define all information that was collected from the patent documents in the patent landscape (e.g. technical area, date of publication), any software used to extract the data, and the protocol if the information sought from a patent document was not available
9	Analysis	Describe any analysis and synthesis of results
10	Patent family designation	State the source of patent family designations (e.g. Derwent or INPADOC) if any analysis incorporated patent families
Results		
11	Patent selection	State the number of patents (or patent families) assessed for eligibility, the number included in the study, and the reasons for exclusion at each stage of the process. A flow diagram may be useful
12	Data standardisation	Provide details of any steps taken to standardise or normalise the data. Examples would typically include correcting

Item # ***Section***	**Topic**	**Checklist Item**
		misspellings, and discussion of assumptions associated with licensing or mergers and acquisitions
13	Summary	Summarise the patents included in the study (e.g. with reference to the data extracted from them, geographical distribution, temporal distribution)
14	Analysis	Present and explain the results of any analysis (statistical or otherwise) conducted. Include details of settings used for any analyses (e.g. spatial concept maps). For any temporal analysis, include details of what year convention was used (e.g. earliest priority year, application year, publication year)
15	List of patent numbers	List the patent publication numbers for any patents included in the study (the supplementary material will often be a suitable location for this)
Discussion		
16	Summary	Summarise the main findings, how they relate to the aims, and to whom they may be relevant
17	Limitations	Discuss any limitations of the work in the context of the reliability of the conclusions; include discussion of limitations related to the methodology and software. If applicable, include information relating to how sources of error were reduced
18	Context	Explain how the findings relate to other studies in the field, how the study builds upon previous work, its potential impact, and implications for future research
19	Conclusions	Provide a conclusion which gives a general interpretation of the results in the context of other evidence
Other		
20	Conflict of Interest	Disclose any potential conflicts of interest
21	Funding	Disclose any sources of funding for the study and the role of the funder in the study, and any other support received during the study (e.g. supply of data)

3.4.6 Definition of a Patent Landscape

Consensus was achieved on the definition of a patent landscape to be the following (adapted from ref: [17] and previously published in ref: [20]):

> *"A landscape is an analysis of the relationships between multiple sets of indicators or of those indicators measured against temporal, technical or spatial dimensions. At least one of the indicators, in the case of a patent landscape, are patent publications or some aspect thereof. A landscape seeks to encompass an entire population of relevant data, rather than a random sample drawn from that population".*

3.5 Discussion

Patent landscapes have been published focussing on a broad range of topics, including inhalers[87], RNA interference[88], antibody therapeutics[89], transgenic cotton[51], and many others. They have been used in attempts to identify innovation drivers[90], predict product development[91], analyse patent practices[64], and assess technological progress[60]. Clearly, therefore, their findings could underpin important decision-making and analysis by academics, industry and government. Despite this, there is evidence that the quality of reporting in patent landscape articles in the life sciences is insufficient.

We noted in the previous chapter that the process of conducting a systematic review is analogous to that of conducting a patent landscape. However, in many cases, patent landscape articles are published as article types other than "original research" and may avoid the close methodological scrutiny that would accompany the publication of a systematic review. For systematic reviews, and many other areas of health research, reporting checklists, flowcharts, or other guidance documents have been developed to improve reporting quality, and there is evidence suggesting that they have done so[44,59,75].

The vast majority of reporting guidelines developed to date focus on health research reporting. Outside of specifically health research, however, the development and implementation of reporting guidelines has been much more limited. It is plausible that the reason for this is related to the direct impact on physician decision-making elicited by published health research, which in turn directly impacts patient care. For example, the findings of a systematic review comparing the efficacy of two interventions may directly influence the choice of therapeutic prescribed by a physician. Therefore, the ability of both the physician and the publisher of the research to evaluate the methodology, results, and other aspects of the paper is of paramount importance, and reporting guidelines to facilitate this have resulted.

Patent landscapes, by contrast, are more influential in the innovation process, and may thus have considerable weight in determining technology development. Largely, decision-making based on patent landscaping will occur at the system-level (such as in strategic decisions in government or companies) and therefore have broad reaching impact. For example, given the UK government's investment in a number of patent landscape reports which, among other things aim to help people "to consider the direction of future funding"[60], it is apparent that patent landscapes are expected to be relied on for such purposes. At the other end of the spectrum, individual researchers or research groups may rely on the findings of patent landscape articles to determine research focus (e.g. if a patent landscape indicates that there are challenges with freedom-to-operate in an area, that area may be avoided by researchers). In both cases, the eventual impact on society could be significant, and, therefore, the ability to evaluate, synthesise and reproduce such work is significant. For these reasons, we have developed a reporting checklist: RIPL.

3.5.1 Scope

The RIPL checklist is intended for patent landscape articles, as defined above, that are published in academic journals. The literature that provided the justification for this checklist development focussed on the life sciences; however, the checklist is not specific to the life sciences and can be applied uniformly regardless of the discipline. In particular, authors and reviewers of patent landscape articles, as well as journal editors who publish them, could benefit from its use. For authors, consulting the checklist prior to and during the development of a manuscript will help to ensure that it is well reported, facilitating fair evaluation of the manuscript by peer reviewers, reproducibility, and improving the feasibility of evidence synthesis. For reviewers and journal editors, the checklist provides a straightforward means to assess compliance to a consensus-based set of minimal criteria that should be reported in a patent landscape article. Journal endorsement of these guidelines could improve the quality of patent landscape articles accepted for publication.

While this checklist is intended primarily for use in academic research, reporting the recommended items in any patent landscape may improve the quality of the report and facilitate its assessment by others. Analysts or researchers in industry, government, or other organisations producing patent landscape reports for internal or external publication may therefore benefit from following the checklist.

It should be stressed that this guidance does not extend to recommendations regarding methodologies; rather, it focusses specifically on reporting. For information on conducting patent landscapes, useful resources are available from the World Intellectual Property Office (WIPO)[92,93]. In particular, when planning a patent landscape, we recommend consulting, in conjunction with this work, the "Guidelines for Preparing Patent Landscape Reports"[92], which provides more granular methodological considerations for conducting patent landscapes, in comparison to this chapter which provides recommended items for reporting. Papers highlighting the need for methodological standardisation may also be useful[17,36].

3.5.2 Explanation of Reporting Items

Reporting items are grouped according to the typical structure of research articles (Table 3.2). We recognise that, in some cases, reporting in this format may not be appropriate and information may need to be included in the supplementary materials or elsewhere in the manuscript. However, this format provides a useful framework for discussing the items. Additionally, recognising that patent landscapes represent original research that others may wish to scrutinise or replicate may improve reporting quality and, therefore, reporting in a structured, research article format may be preferable when possible. In our systematic review (Chapter 2), research articles were reported marginally but significantly better than review articles. Here, we briefly explain the rationale for each item in the checklist. Where there is overlap with items in other guidance, similar rationale is provided here[33,94–96].

3.5.2.1 Title

Item 1, Title: Identify that the article includes a patent landscape and state the subject matter under investigation (e.g. gene editing technologies)

Information or insight in patent landscapes can only be used if the papers reporting them can be identified. Given the increasing number of research articles published[11], identifying relevant literature is challenging and lack of ability to identify papers may lead to wastage and duplication of effort. Stating that an article is a patent landscape in the title is therefore important. For the same reasons, the subject matter should be included in the title. Delphi participants did not think that the main conclusion of the article should be stated in the title, possibly mirroring other concerns that doing so can exaggerate findings[33].

3.5.2.2 Summary/Abstract

Item 2, Abstract: Provide a summary which includes the background, rationale, results and main findings in the context of the aims

Abstracts provide important information that allows the reader to determine whether the full article should be read, and, with the title, may be the only information available to readers[33] (e.g. if articles are behind a paywall). As a minimum, the reader should be able to understand the background and rationale of the study, the results and the main findings in the context of the aims. If journal policy permits, including methodological information is also useful; however, key journals publishing patent landscapes permit only short abstracts for patent articles (e.g. [15,97]) so this may not always be possible.

3.5.2.3 Introduction

Item 3, Rationale: Describe the rationale for the study, including relevant background information and the potential impact of the investigation

Patent landscapes may be conducted for a number of reasons already outlined. Therefore, it is important to understand the specific rationale for the reported study to allow readers to determine relevance. This would typically include briefly discussing current evidence or work in the field and its limitations, along with what the patent landscape hopes to add to this work and why that is useful.

Item 4, Aims: Describe the aims of the study

Providing an explicit statement of the question being addressed by the patent landscape is essential. The aims should allow the reader to understand the article's scope and therefore applicability. Explicit aims are particularly important because they assist in determination of methodological elements[33] (e.g. patent eligibility criteria and analysis) and facilitate critical appraisal of the study[95]. Stating that the aim was "to review the patent landscape of X" is not sufficient because the term "patent landscape" does not represent a single methodology and patent landscapes can be conducted for many reasons. A statement describing what the patent landscape aims to achieve is more useful, such as: "Our objective in this paper is to shed light on the use of primary and secondary patents by multinational originator companies in Chile and to gauge their effect on creating and maintaining exclusivity"[98]. In some cases, it may be useful to report aspects of Item 3 with the aims, such as in the following example: "The information resulting from this study may be of value to both R&D managers and researchers in the field by identifying the most relevant technologies and giving an overview of the rFVIII inventions in the last 20 years”[97].

3.5.2.4 Methods

Item 5, Search: State the databases and patent offices searched, the dates on which searches were conducted, and the components of the patents searched. Include the search terms used for all databases searched

Depending on the aims of the patent landscape and availability of resources, different databases and patent offices may be searched. Searching different databases and patent offices leads to different results (e.g. European patents vs. US patents) so the details must be disclosed for the purposes of reproducibility and determination of applicability. Similarly, the dates for searches (including any prior date restrictions) and components of patents searched (e.g. title, abstract, whole text, inventors, classifications) influence the retrieved patents and must be reported. Developing search terms is a time consuming process requiring subject matter expert input[17], and search terms, therefore, represent a valuable resource. Reporting the full search terms, including specifying for which databases they were used, allows searches to be re-run simply. This is useful for critically appraising articles, reproducing findings, and updating landscapes. If possible, the dates and components of the patents searched should be incorporated into the search terms. Using patent classification codes, such as international patent classification (IPC) or cooperative patent classification (CPC) codes, to narrow search results is often useful; details should again be included in the reported search terms. Supplementary materials are often an appropriate location for search terms given their complexity and potential size. Useful guidance on quality control for patent search strategies is available[99].

Item 6, Selection criteria: Include details of the selection criteria of patents to be included in the patent landscape, including the subject matter of those patents

There are a large number of published patents, and the perceived relevance of a patent to a particular research question could vary between researchers. Reporting the selection criteria for patents included in the landscape is, therefore, important for understanding the landscape's comprehensiveness, validity and applicability. If any sorting of patent documents is conducted the selection criteria are also essential for reproducibility. Selection criteria should include information such as: language restrictions, jurisdiction(s) of interest, patent status (e.g. applications, grants, all), and, importantly, subject matter. Subject matter should include the general field of interest (e.g. small molecule therapeutics for melanoma) but also, if applicable, which patents within that area are relevant (e.g. product, manufacturing, all).

Item 7, Identification of relevant patents: If applicable, state how patents identified in searches were sorted for relevance

Patent search results often require sorting to remove irrelevant documents, either computationally, through manual review, or both. If applicable, details of this process should be reported to allow reproducibility and assessment of the likely levels of noise in the data. Reported information should include which parts of the patents were used in sorting for relevance (e.g. title, abstract and claims were reviewed to exclude irrelevant patents), who conducted the sorting or which software was used, and any measures taken to enhance objectivity, such as repetition of part or all of the process by an independent investigator. If input from experts, such as subject matter experts or patent agents was sought, this should be stated. If identified patents were not sorted, for example because the number of documents was impractically large, this should be reported.

Item 8, Data extraction: List and define all information that was collected from the patent documents in the patent landscape (e.g. technical area, date of publication), any software used to extract the data, and the protocol if the information sought from a patent document was not available

In almost all forms of patent landscape, some information is extracted from patents and defining it is essential to understand later analysis. Examples include the date of patent publication, the technical area of the patent, inventor or assignee names, patent classification codes and number of claims. Details of the information extracted from each patent should be reported and if software was used for extraction, information such as name and version should be given. If an information extraction sheet was used, providing it may be useful. As in Item 7, if applicable, any measures to enhance objectivity and input from experts should be reported. Information sought from patents may not be available from every patent; the protocol employed in this scenario should be stated as treatment of missing data, or assumptions relating to it, might influence results.

Item 9, Analysis: Describe any analysis and synthesis of results

All patent landscapes, by definition, include analysis of patent documents. This analysis provides the central means by which the aims of the landscape are addressed and is critical to report. Examples of analyses include counts of patents applications or grants over time, citation network analysis, assignee network analysis, thematic clustering, geographical analysis and text-mining[97,100,101], all of which differ in the specific procedures. Details of the methods for analysis and synthesis of patent documents should be given that are sufficient to allow reproducibility. This should include details of the software and software version used, any assumptions made about the data, any statistical analysis conducted, and which information was included in which analysis (for example, a subset of patent documents may be used for certain analyses). Preferably, analysis should be planned in advance[36], and noting which analyses were planned post-data collection would be useful.

This item has overlap with Item 12 (data standardisation) in the results section, details of which could optionally be reported in the methods section. Reporting data standardisation in the methods section may be more appropriate if uncommon methods which require explanation are used, such as custom programming scripts. Some aspects of Item 14 (analysis) might also be reported in this section, depending on the details of the analysis. For example, if multiple spatial concept maps are generated, each with different settings, reporting the specific settings alongside the results might be preferred; if multiple thematic spatial concept maps are generated using the same settings, reporting those settings once in the methods may be more efficient.

Item 10, Patent family designation: State the source of patent family designations (e.g. DWPI or INPADOC) if any analysis incorporated patent families

It is often useful to analyse patent families as well as or instead of individual patent documents. Patent families generally represent distinct inventions, whereas there can be many patent documents associated with a single invention (e.g. multiple national filings). However, there are multiple definitions of patent families and different methods available for generating them, all of which will influence results of analysis and are associated with different limitations[92]. Common patent family designations include Derwent World Patent Index (DWPI), International Patent Documentation (INPADOC), or European Patent Office simple patent families, but other options are also available. If patent families are used in analysis, their source and definition should be reported.

3.5.2.5 Results

Item 11, Patent selection: State the number of patents (or patent families) assessed for eligibility, the number included in the study, and the reasons for exclusion at each stage of the process. A flow diagram may be useful

It is helpful for readers to understand how the data collection process resulted in the final patent documents included in the landscape. As a minimum, the initial number of patents returned from searches, the final number of patents included in the landscape, and the reasons for exclusion should be reported (if applicable). Reporting reasons for exclusion is important to assess how the selection criteria were applied. Ideally, the total number of patents retrieved from each search (e.g. of different patent offices or databases) should be stated separately and the effect of combining, collapsing and cleaning the data on the number of documents provided. Flow diagrams are often recommended for similar dataflows in other fields[33], and may be useful to summarise the steps taken between original searches and arrival at the final dataset(s).

Item 12, Data standardisation: Provide details of any steps taken to standardise or normalise the data. Examples would typically include correcting misspellings, and discussion of assumptions associated with licensing or mergers and acquisitions

Raw patent data is often "messy" and may require cleaning before analysis[92]. For example, if analysing patent assignees, a single commercial entity may be represented by different suffixes (e.g. Inc. vs. Ltd.), or the same suffix presented differently (e.g. Limited vs. Ltd.) and it may be necessary to combine these. There are often misspellings in assignee and inventor names which require cleaning for accurate results. Mergers and acquisitions may need to be considered and acquired companies combined with their now parent companies. These cleaning steps can be done using automated or manual methods, and details should be provided. As stated in the explanation for Item 9, this information may alternatively be reported in the methods section.

Item 13, Summary: Summarise the patents included in the study (e.g. with reference to the data extracted from them, geographical distribution, temporal distribution)

A summary of the patent dataset is needed to understand its scope and comprehensiveness. As well as number of patents (Item 11), other useful and common summaries include geographical distribution, temporal distribution (noting details as per Item 14) and summarisation of any data extracted from patents (e.g. number of process vs. number of product patents).

Item 14, Analysis: Present and explain the results of any analysis (statistical or otherwise) conducted. Include details of settings used for any analyses (e.g. spatial concept maps). For any temporal analysis, include details of what year convention was used (e.g. earliest priority year, application year, publication year)

The results should be presented and explained such that any decisions made with respect to treatment of the data and details of the analysis are transparent. Many automated analyses have a large variety of settings for customisation, and details of these should be included so that the analysis could be reproduced with the same data. For spatial concept maps, any manual adjustment, such as grouping or exclusion of terms, should be described. For temporal analysis, it is important to state which year convention is used, as year conventions influence results and their interpretation. If different data are used for different analyses, this should be stated and the differences in the data explained. Whether patent documents or families are used for analysis should be stated.

Item 15, List of patent numbers: List the patent publication numbers for any patents included in the study (the supplementary material will often be a suitable location for this)

The patent publications included in the landscape are the data for analysis and underpin the findings of the study. As all of the documents will be available in the public domain, providing a list of the patent numbers included in analysis allows other researchers to repeat or check any analysis without the need to repeat searches and time-consuming sorting steps, and also to directly assess the relevance of data to their interests. In most cases, supplementary material will be a suitable location for this information. Some landscapes include very large numbers of patents so providing a link to a data repository with the information may be preferred.

3.5.2.6 Discussion

Item 16, Summary: Summarise the main findings, how they relate to the aims, and to whom they may be relevant

A summary of the main findings, how they relate to the original aims, and the applicability of the findings to different stakeholders should be provided, as is common in most research.

Item 17, Limitations: Discuss any limitations of the work in the context of the reliability of the conclusions; include discussion of limitations related to the methodology and software. If applicable, include information relating to how sources of error were reduced

All patent landscape studies are likely to have limitations, and these should be discussed. Methodological limitations can include those related to the data or to the analysis. Often, relatively simplistic analyses are performed which aim to address complex questions. For example, spatial concept maps may be used to identify gaps in technology development. If readers are not closely familiar with the patent landscaping process and the methods for generating spatial concept maps, “gaps” may be misinterpreted as areas where there is no patent activity. However, without specific analysis of claims such inference cannot be made. Similarly, counts of patents are a common form of analysis but are known to provide little insight[49]. Limitations in analyses should therefore be discussed so the reader is not misled. Discussing limitations in the data is important to allow assessment of the generalisability of the findings. For example, geographical or language restrictions, or changes in institutional regimes that may influence the observed trends[36], would be useful to highlight. Software used in analysis of patent documents also has limitations which should be discussed: for example, automated data cleaning may be imperfect. If sources of error were reduced, for example by manual cleaning, this information should be included.

Item 18, Context: Explain how the findings relate to other studies in the field, how the study builds upon previous work, its potential impact, and implications for future research

If previous work has been conducted in the same or a similar area, a more detailed discussion of the differences between that work and the authors' work is useful at this stage. This should include how the work builds on previous studies, what the impact of the study is and what implications might be for future research.

Item 19, Conclusions: Provide a conclusion which gives a general interpretation of the results in the context of other evidence

A conclusion summarising and providing a general interpretation of the results in the context of the work discussed in Item 18 should be provided.

3.5.2.7 Other

Item 20, Conflicts of interest: Disclose any potential conflicts of interest

The subject matter of patents is inherently of commercial interest. It is therefore essential that any potential conflicts of interest are disclosed so that readers can assess likelihood of bias. Conflicts of interest should be reported in accordance with journal policy. If no policy is available, guidance is available for medical research reporting that could be used in this context[102].

Item 21, Funding: Disclose any sources of funding for the study and the role of the funder in the study, and any other support received during the study (e.g. supply of data)

Commercial funding has been associated with increased likelihood of reporting favourable results in other research areas[103,104]. Disclosing sources of funding is therefore important to allow assessment of credibility and bias, particularly because patent landscapes may be concerned with strategic questions with commercial value. For the same reasons, the role of the funder and any support received during the study should also be reported.

3.5.3 Limitations

Numerous methods have been used to develop reporting guidelines, though it is generally agreed that a consensus process should be incorporated as a central element of the methodology[86]. The only guidance available for development of reporting guidelines assumes that a face-to-face meeting is conducted[105]; however, nearly 50% of guideline development processes do not include one[86]. Given the desire to include international participants in our consensus process, and the extensive additional resources required to conduct face-to-face meetings, we judged that a Delphi study, followed by finalisation of the checklist by the authors, represented the most reasonable methodology. It could be argued that some opportunity for discussion was lost because of this. Using a Delphi study as the method for achieving consensus, however, allows anonymity and confidentiality to be maintained, reducing known negative characteristics of group dynamics such as the influence of dominant individuals[106].

The outcome of a modified Delphi study, and any other consensus methodology, is influenced to some degree by the quality and expertise of the participants. We identified experts through a number of methods, some of which could introduce bias to the study (e.g. authors recommending participants for the study). Though such sampling is not ideal, it is common practice in the development of reporting guidelines and it was felt necessary to ensure a sufficient sample size was recruited. To mitigate any bias that may have been introduced in this way, we welcome further comment and feedback on the reporting checklist. As practices evolve over time, or indeed become better understood, so too may the requirements for reporting. Updates to this checklist may, therefore, be conducted in the light of feedback received from the readership of the paper resulting from this chapter.

3.5.4 Endorsement

Journals publishing patent landscape articles may benefit from endorsing and encouraging adherence to the proposed checklist for reasons already outlined. A uniform endorsement policy across journals may help to ensure optimal uptake and ensure that the requirements for authors are the same across multiple outlets, thus reducing the burden of compliance. We therefore propose that journals publishing patent landscapes adopt the following text in endorsing this guideline, for example in the "Instructions for Authors" (adapted from the PRISMA-P 2015 statement[107]):

"[This journal/organisation] endorses reporting according to the Reporting Items for Patent Landscapes (RIPL) checklist. We recommend that, while preparing an article incorporating a patent landscape, all items in the RIPL checklist that are applicable to your study are reported. Ensuring that these minimal items are reported will improve the manuscript and potentially improve its chances of eventual acceptance."

3.5.5 Conclusion

Reporting checklists have been widely developed and deployed in health research, though outside of specifically health research are far less prevalent. Patents are a rich source of information that can be analysed for a variety of purposes *via* patent landscapes, and the resulting publications may be relied on for significant decisions. The checklist presented aims to improve the quality of reporting in such publications, but also highlights the importance of reporting quality beyond those areas in which it has traditionally been a focus. Through the implementation, evaluation and continuous improvement of these recommendations, all stakeholders associated with patent landscaping, from authors to reviewers and editors to innovators, could benefit.

4 Quality and Performance of Algorithms for Predicting Drug Approval: A Systematic Review

A protocol for this systematic review is registered on PROSPERO (www.crd.york.ac.uk/prospero/) with ID: CRD42018093735. Citation: "Smith JA, Van Velthoven M, Halioua C, Meinart E, Carr AJ, Brindley DB. Performance and quality of algorithms for prediction of therapeutic market authorisation: systematic review protocol. PROSPERO, CRD42018093735, 2018."

4.1 Abstract

High attrition during drug development is a major driver of the cost of new medicinal drugs. The ability to predict or prioritise drug candidates that are eventually approved could therefore be of considerable value, though at present, it is not clear what attempts to do so exist and how well they perform. We therefore conducted a systematic review of papers developing or validating multivariable computational algorithms for the prediction of drug approval, with the aim of describing their methodologies and assessing their quality and performance. Three studies representing four distinct models were identified. Reported events per variable were very low (less than five) in all models, and two of the papers exhibited further methodological issues that are likely to result in very biased models. One paper used generally appropriate methodology and reported potentially useful predictive performance (~0.8 area under the receiver operating characteristics curve [AUC] from the end of phase 2 and phase 3) but did not report the final model in a usable way. A brief review of other relevant drug success

literature is also provided. In conclusion, there is limited utility to any existing published predictive models of regulatory approval, though there is evidence that prediction of approval is feasible to some extent.

4.2 Introduction

The consistently increasing cost of bringing new drugs to market is so pronounced that it has been dubbed “Eroom’s Law” (Moore’s Law, backwards)[108]; the capitalised cost of bringing a new molecular entity (NME) to market is estimated to be as high as $2.6 billion[3]. A major driver of this cost is the attrition rate during drug development. Estimates vary, though it is generally believed that only around 10-15% of drugs that enter first-in-human trials are eventually approved[109–111]. Not only is this an inefficient use of capital, but it also results in patients receiving treatments in clinical trials that ultimately prove ineffective. Therefore, efforts to predict which drug candidates are likely to be successful could be of considerable value, both in improved capital efficiency, and improved and expedited outcomes for patients.

In vitro and *in vivo* experimentation has traditionally been, and remains, one of the key drivers of decision-making for early stage compounds. However, flawed preclinical research[112] and inherent limitations in extrapolation from animal models to humans[113] have led many to doubt the suitability of these models alone to predict which molecules will eventually succeed. *In vitro* and *in vivo* research is also time consuming and costly.

In parallel to traditional laboratory-based experimental approaches, therefore, a number of rules to assist in the identification of promising molecules in drug discovery have arisen. A classic example of this is Lipinski’s “Rule of 5”, which predicts poor absorption or permeation for oral drugs when several physicochemical parameters are outside of certain ranges[114] (molecular weight ≤ 500, logP ≤ 5, H-bond donors ≤ 5, H-bond acceptors ≤ 10). Compounds within these ranges are considered more "drug-like" than those outside of them. Since Lipinski's publication, the concept of drug-likeness has developed considerably, and different rule-based criteria have been proposed for drugs delivered *via* other means or used for specific indications, and have been extended outside of specifically drugs to leads and tools[115]. They have been developed in numerous ways, many relying on analysis of historical data using counting, machine learning, statistical models and other computational approaches and have been extensively reviewed[116–119].

Much less common are efforts to predict directly, from either discovery or development stages, which molecules will eventually receive market approval: the ultimate aim of any drug research and development program. With improved computational power and approaches such as machine learning, attempts to do so may be becoming more feasible, and improved availability of relevant historical data may compound this.

However, at present, to our knowledge, no systematic synthesis of the attempts made through computational methods to predict marketing approval exist. Such a synthesis would be useful for several reasons: i) identification of models available for the benefit of potential users of those models; ii) comparison and critical appraisal of the methods used in developing and validating models to evaluate their relative performance; iii) identification of promising characteristics of existing approaches, such as aspects of the design, analysis, or predictor variables, for integration and further development in future approaches.

For these reasons, this chapter details a systematic review for identification, description, and assessment of multivariable computational algorithms to predict marketing approval of therapeutic candidates at any time point during discovery and development prior to market authorisation. Note that the terms model and algorithm are used in interchangeably in this chapter. Specifically, the systematic review has the following aims:

1. Identifying studies developing algorithms for prediction of marketing approval of therapeutic candidates

2. Describing the methodological approaches taken to develop the algorithms and describing the algorithms themselves

3. Assessing the performance of those algorithms in light of their methodological quality.

4.3 Methods

A protocol for this systematic review was registered on PROSPERO[21]. This systematic review is reported, where applicable, in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist[34]. The Cochrane Collaboration and Centre for Review and Dissemination (CRD) methodology for conducting systematic reviews is followed where appropriate[120,121].

4.3.1 Eligibility Criteria

The Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist was used to inform the development of the eligibility criteria[122]. Though it is aimed largely at clinical prediction studies, many of the considerations raised are applicable.

4.3.1.1 Study Design and Scope

Studies developing or validating computational algorithms to predict approval of therapeutic candidates were included. Studies had to include at least two groups, of which one is approved drugs, and another is abandoned drug candidates no longer under development. In addition, an attempt to differentiate between those two groups based on data that would be available before approval was required (i.e. the study had to be prognostic rather than diagnostic). Only algorithms developed based on computational analysis of multiple predictor variables were included. Qualitative studies, or studies incorporating patients as the primary unit of analysis rather than drugs, were not eligible for inclusion: for example, clinical studies.

4.3.1.2 Population to which Predictive Model Applies

Studies analysing multiple failed drug candidates and multiple approved drugs were included. There were no limitations with respect to the types of drugs included in the models: models developed using biological and/or small molecule drugs were applicable. Studies focussed on devices, diagnostics, or other non-drug medical interventions were excluded.

4.3.1.3 Type of Prediction Modelling Study

Studies developing and/or validating prognostic models were included.

4.3.1.4 Outcome to be Predicted

Studies had to aim to predict regulatory approval (also known as marketing approval, regulatory success, market authorisation and others) and failure during drug development. This failure could occur any time prior to market approval. Prediction of failure could be implicit; for example, the prediction of a low probability of approval.

4.3.1.5 Intended Moment of the Model

The intended moment of the model had to be any point prior to approval. For example, studies aiming to predict approval from preclinical stages or clinical phases would both be eligible. Any timespan of prediction is therefore appropriate.

4.3.1.6 Comparators

Comparators in the context of this review were defined as methods or techniques against which the performance of the algorithm developed is compared: for example, other models to predict approval, or expert opinion. Studies with any or no comparators were eligible for inclusion.

4.3.1.7 Report Characteristics

Only English language articles were eligible for inclusion due to resource constraints. Articles published at the time of conducting the search and pre-publication prints, if available online, were included. Summaries or abstracts from conferences were excluded.

4.3.2 Information Sources

Electronic database searching was the primary means of identifying relevant articles. Articles meeting our eligibility criteria could potentially be published in journals indexed in a number of different subject areas, depending, for example, on the variables used to develop the algorithm. Therefore, scoping searches were conducted in the Scopus database, which includes titles in the following four areas in which most, if not all, relevant articles would likely fall: life sciences, physical sciences, health sciences, and social sciences & humanities. We also searched PubMed, EMBASE (1974 to present) and MEDLINE (the latter two *via* the Ovid interface), three databases commonly used for biomedical literature searching.

Authors of potentially relevant studies were contacted to request full text articles if the full text article could not be freely accessed by one of the authors of our systematic review.

The electronic database search was supplemented by screening of the references of articles identified as relevant. Grey literature was not systematically searched, though if relevant material was identified it was noted and referred to.

4.3.3 Search Strategy

We developed a search strategy for use in Scopus which was published in the systematic review protocol[21] and which was not modified (Supplementary Table 8.5). The search strategy was developed by the authors of the systematic review protocol who have previous experience in conducting systematic reviews, including in the development of search terms[20,123]. Search strategies for PubMed, MEDLINE and EMBASE were subsequently developed (Supplementary Table 8.5, Supplementary Table 8.6). There were no prior date restrictions on any of the searches and dates of search are provided in the relevant tables.

4.3.4 Data Management and Article Selection

Search results were imported into RefWorks (www.refworks.com), where duplicate entries were removed using an automatic function and through manual screening of closely related articles. Titles and, where available, abstracts of the remaining articles were screened against the eligibility criteria by two independent reviewers; if it was unclear based on the title and abstract whether a study should be included, it was not excluded at this stage. The remaining articles were screened as full text articles against the inclusion criteria, and the reasons for exclusion from this stage onwards were recorded. Full text articles that could be accessed freely by one of the authors of our systematic review were included; if the full text article was not accessible, we agreed to email an author with a request to provide the full text article. If the full text article was still not available, we agreed to record the article as potentially relevant. No article information was blinded or masked from the reviewers.

The selection process was piloted by two reviewers on a random sample of 10% of retrieved articles of the first database searched. The article order was randomised for each reviewer. Differences in results were discussed by the authors to determine whether they should be attributed to inadequately defined inclusion criteria or to misunderstanding of appropriately defined criteria. It was agreed that if the pilot demonstrated that the inclusion criteria were inadequate for consistent application, they would be refined and the pilot study repeated with a new random sample of the retrieved articles. This process would be repeated until consensus was achieved, at which point the reviewers would continue to assess the remaining articles, and those retrieved from searches of additional databases. Consensus was defined as ≥95% agreement with respect to inclusion or exclusion of articles, following discussion to resolve differences.

The independent reviewers compared and discussed discrepancies following the title and abstract screening stage to agree on a single set of full text articles to review. When we disagreed over the inclusion of an article at this stage, it was included. In the subsequent full text review, any differences in opinion relating to inclusion or exclusion of articles were resolved through discussion between the reviewers or, if unsuccessful, consultation with a third reviewer.

The same study or studies examining the same dataset may be reported in several articles. To determine if this was the case, we compared the author lists of identified papers for overlap and compared the summary statistics of reported datasets. In the case that two articles reported the same model and met the inclusion criteria, both studies were included for data extraction, but at that stage were consolidated and treated as though reported in a single article.

4.3.5 Data Extraction

Standardised data extraction forms were used to extract data from relevant studies. These were based on items from the CHARMS checklist[122], adapted to be suitable for non-clinical modelling (Supplementary Table 8.7). We planned to pilot a draft of the data extraction form to ensure relevant information was consistently captured. Because so few papers were identified (see Results), a second author instead checked all data extraction. Not all data detailed in the data extraction form was expected to be relevant to all articles; where the data was not reported, this was noted.

4.3.6 Quality Assessment

The extracted data were compared to the methodological recommendations in the CHARMS checklist[122] and Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement[124] and explanation and elaboration[95] (TRIPOD provides more detailed recommendations but is aimed at model developers). We focussed on items that might influence the risk of bias for each study. Risk of bias refers to “the extent that flaws in the design, conduct, and analysis of the primary prediction modelling study lead to biased, often overly optimistic, estimates of predictive performance measures”[122]. Both documents focus on clinical prediction modelling but the methodological considerations should apply more generally. Because so few papers were identified, only qualitative data synthesis and analysis are provided.

4.4 Results

4.4.1 Search Results and Study Inclusion

We identified three papers relevant for inclusion in this review (Figure 4.1). The pilot study indicated that the article selection process was acceptable. In the full study, following duplicate removal, the titles and abstracts of 1,617 articles were screened for relevance, and 1,564 articles were excluded. The remaining 53 articles were assessed as full-texts against the eligibility criteria and 51 were excluded: seventeen did not present a computational algorithm, 32 did not predict approval, 19 did not develop a model from a dataset of approved and discontinued drugs, five did not use multiple predictor variables, and 17 were the incorrect study type (note that each article was excluded for at least one reason, but the reasons for exclusion are not necessarily exhaustive for each article). Two of the included articles were identified through this search method and were published in journals[125,126]. An additional article for inclusion was identified[127] through general internet searching (Google) using key words from the database search strings. The additional article was a pre-print or working paper, and therefore may not represent the final work. This should be considered in the following analyses.

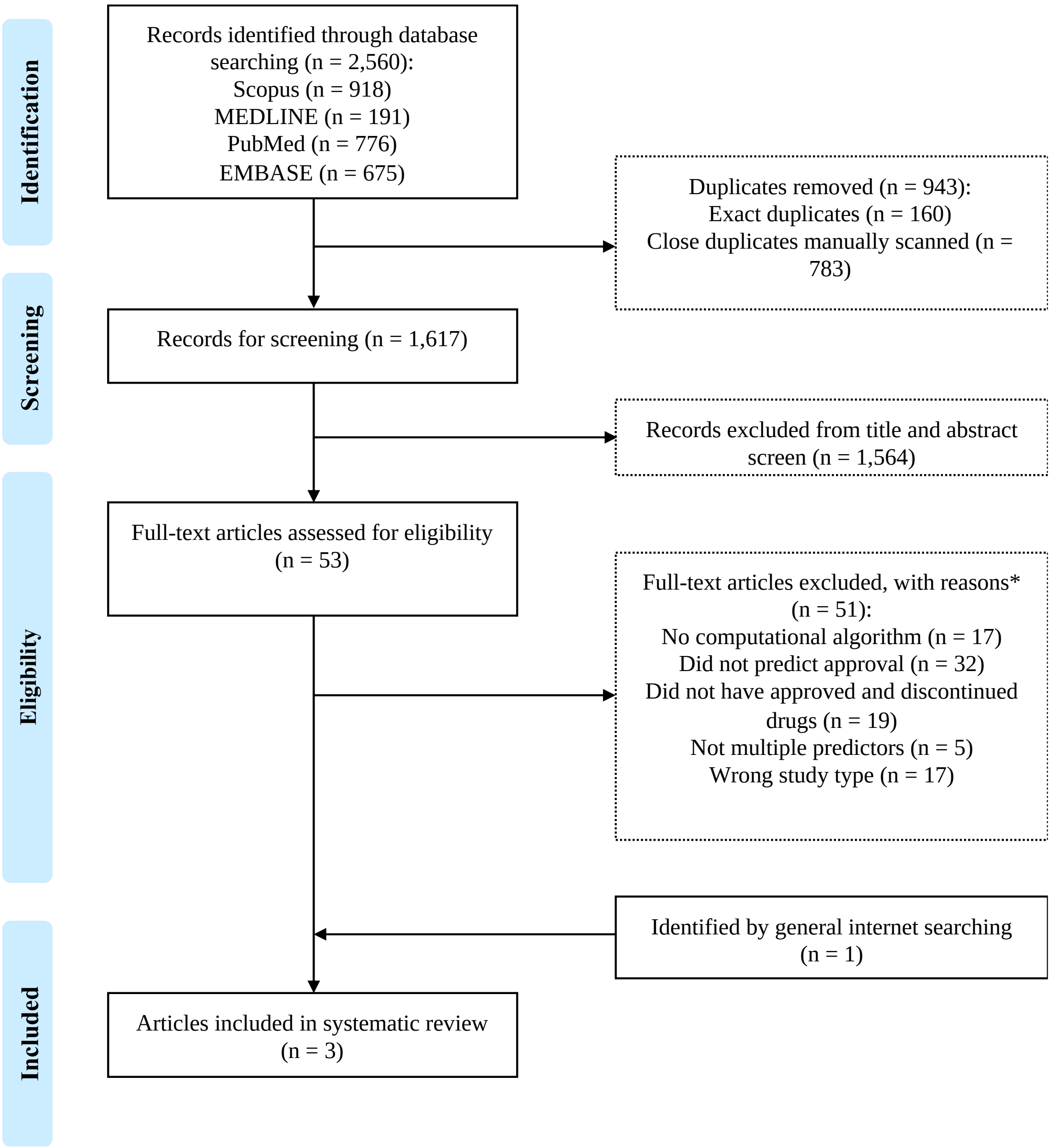

Identification
Screening
Eligibility
Included
Records identified through database searching (n = 2,560):
Scopus (n = 918)
MEDLINE (n = 191)
PubMed (n = 776)
EMBASE (n = 675)
Duplicates removed (n = 943):
Exact duplicates (n = 160)
Close duplicates manually scanned (n = 783)
Records for screening (n = 1,617)
Records excluded from title and abstract screen (n = 1,564)
Full-text articles assessed for eligibility (n = 53)
Full-text articles excluded, with reasons* (n = 51):
No computational algorithm (n = 17)
Did not predict approval (n = 32)
Did not have approved and discontinued drugs (n = 19)
Not multiple predictors (n = 5)
Wrong study type (n = 17)
Identified by general internet searching (n = 1)
Articles included in systematic review (n = 3)

Figure 4.1: PRISMA flow of articles

PRISMA flow diagram[34] *detailing number of studies included at each stage and reasons for removal. *Full text articles were excluded for at least one reason, so the sum of individual reasons is not equal to the number of articles excluded.*

4.4.2 Dataset Characteristics

The characteristics of the datasets used to develop the predictive models are shown in Table 4.1. All articles include both large (biological) and small molecule compounds, with DiMasi *et al.* (2015) and Heinemann *et al.* (2016) focussing exclusively on oncology[125,126] and Lo *et al.* (2017) including a wide range of indications[127]. All models aim to predict approval, though the geographical location of approval varied. DiMasi *et al.* develop one primary model, which aims to predict approval from completion of phase II testing (the "DiMasi model"). Lo *et al.* develop two models for predicting approval: one from completion of phase II testing (the "Lo P2APP model") and one from completion of phase III testing (the "Lo P3APP model") from two different datasets. In our results tables, differences between the models are stated. Heinemann *et al.* do not specify a particular time point for prediction; rather, they predict approval from a number of specific time points prior to the outcome (approval or failure), therefore generating a different model for each time point. Detail is not provided on each individual model and the methods are the same for the different models, so we refer to them collectively (the "Heinemann model").

4.4.2.1 Sample Size and Events per Variable

Sample sizes vary greatly across models, as do the number of candidate predictor variables (Table 4.1; detailed discussion of candidate predictors in Section 4.4.2.2). In all cases, events per variable (EPV) are much lower than recommended (at least 10 are usually recommended, but often many more)[95,122,128]. The DiMasi *et al.* dataset comprises 98 drugs; however, in several cases it is not clear how many of these drugs are used in each analysis. Their final scoring algorithm is developed on 62 compounds, though methods used to inform the predictors used in this final algorithm may have used more or fewer compounds. In calculating EPV, we therefore make the optimistic assumption that all events are included in the development data (21 approved drugs). Twenty-three candidate predictors are presented, resulting in an EPV of 0.9. Lo *et al.* use considerably larger datasets (Lo P2APP, 4,812 drug-indication pairs, 13.2% approved; Lo P3APP, 1,610 drug-indication pairs, 40.9% approved) but include a much larger number of candidate predictors (at least 147), leading to maximum EPV of 4.3 and 4.5 respectively. The Heinemann model contains 42 approved drugs and lists nine predictor “classes”, which are further subdivided into a total of 33 predictors. It is assumed that all 33 are considered as candidate predictors, leading to an EPV of 1.3.

4.4.2.2 Details of Candidate Predictors

A wide range of predictor variables are used across the three papers. DiMasi *et al.* list 23 candidate predictor variables for which data was collected across the following groups: drug characteristics, commercial and company, trial design, and trial outcomes. There is considerable overlap with Lo *et al.*, who list 147 example predictors divided into drug and trial predictor variables. However, Lo *et al.* also state “transforming multi-label parent predictors into binary child predictors (1 or 0), there were over 3,000 drug and trial categories in total” and that near-zero variance predictors were removed. The exact number of predictors considered for modelling is not stated. Heinemann *et al.* use a very different collection of predictors, focussing on academic publication patterns for target-indication pairs of drugs. These can broadly be classified as publication counts, information on authors, article indexing terms, and keyword term counts.

Table 4.1: Dataset characteristics of included papers

Item	**DiMasi *et al.*, 2015**	**Lo *et al.*, 2017 (P2APP)***	**Lo *et al.*, 2017 (P3APP)***	**Heinemann *et al.*, 2016**[†]
Drug population	Large or small compounds from top 50 biopharma companies	Drug-indication pairs including biologics and small molecules		Targeted biologics and targeted small molecules
Source of data	Public domain and Tufts Center for the Study of Drug Development	Pharmaprojects and Trialtrove (Informa®, London, UK)		Approved: National Cancer Institute website and FDA website. Failed: Pharmaprojects and Trialtrove (Informa, London, UK), clinicaltrials.gov, Publication information: MEDLINE using the text mining system I2E
Date ranges for included drugs	First entered clinical testing from 1997 - 2007	Aug 8th, 1990 - Dec 15th, 2015	Jan 1st, 1988 - Nov 1st, 2015	Approved drugs (1995 - 2014), failed drugs (2001 - 2014)
Indications covered	Oncology	Alimentary, anti-infective, anti-parasitic, blood and clotting, cardiovascular, dermatological, genitourinary, hormonal, immunological, musculoskeletal, neurological, anti-cancer, rare diseases, respiratory, and sensory products.		Oncology
Positive outcome definition	Approval in the United States, Europe, or Japan	Approval in any market		FDA approval
Negative outcome definition	Termination after phase II or III testing	Suspension, termination, or lack of development		Failure in PII/III clinical trials

Item	**DiMasi *et al.*, 2015**	**Lo *et al.*, 2017 (P2APP)***	**Lo *et al.*, 2017 (P3APP)***	**Heinemann *et al.*, 2016**[†]
Moment of prediction	Completion of phase II	Completion of phase II	Completion of phase III	Various time points defined by years before outcome event
Dataset size	98 drugs (21.4% approved)	4,812 drug-indication pairs (13.2 % approved)	1,610 drug-indication pairs (40.9% approved)	116 drugs (36.2% approved)
Candidate predictor variables	23 candidate predictors in the following groups: drug characteristics, commercial and company, trial design, trial outcomes	147 example predictors listed, divided into drug and trial predictors. Does not appear to be the full list of candidate predictors		9 predictor classes listed with 33 predictor variables in total, all related to academic literature
Events per variable (EPV)[‡]	0.9	4.3	4.5	1.3
% of entries with missing data, by outcome if available	85%	For unique drugs: 36% approved, 71% failed. For unique trials: 78% approved, 40% failed	Unclear (reported only for predictor groups or graphically)	Not specified but likely to be zero given data collection methods

** Lo et al. present two models meeting our inclusion criteria, each making predictions from a different moment of prediction. P2APP = phase 2 to approval, P3APP = phase 3 to approval.*

† Heinemann et al. do not specify a particular time point for prediction, rather, they predict approval from a number of specific time points prior to the outcome (approval or failure), therefore generating a different model for each time point. Detail is not provided on each individual model, so we refer to them collectively

‡ The number of events corresponds to the number of occurrences of the rarer of the two outcomes (approval in all cases). The number of variables is the number of candidate predictor variables considered for input into the multivariable prediction model. Highest plausible EPV presented.

4.4.3 Missing Data

Reported levels (Table 4.1) and treatment (Table 4.2) of missing data across the three studies differed considerably. DiMasi *et al.* report that 15 of 98 observations had complete data. Complete case-analysis is used, though the number of complete-cases varies because different combinations of predictor variables are used for different analyses. There were 62 compounds with all predictors selected for use in the final model. This approach is both inefficient and likely to lead to biased results[95]. Heinemann *et al.* do not report any missing data and based on their data collection methodology it is likely that there was none. Lo *et al.* report graphically and in tabular form the levels of missing data, though number of observations with missing data cannot be extracted for P3APP. For P2APP, levels of missing data are provided for unique trials and unique drugs, but not in an overall summary showing missingness per drug-indication pair (the unit used for prediction). However, they use a suitable approach for dealing with the missing data, evaluating a number of methods experimentally (listwise deletion, unconditional mean imputation, k-nearest neighbour imputation [k-NN], multiple imputation, and decision tree algorithms), and select 5-NN imputation as the preferred approach.

4.4.4 Modelling Approaches

Across all three papers, model development consists of an initial stage in which several machine learning or regression approaches are evaluated to select an approach for the final model (Table 4.2). Both Lo *et al.* and Heinemann *et al.* use the random forests algorithm for the final model, and do not appear to use predictor selection methods before or during modelling, apart from removal of near-zero variance predictors by Lo *et al.* Both models also treat continuous variables as continuous, which is generally recommended[95]. DiMasi *et al.*, by contrast, use machine learning methods, logistic regression and univariate analysis to identify important predictor variables and optimal cut-points for continuous variables, and use the results to develop a simple scoring algorithm that can be used manually. The scoring system comprises four originally continuous variables, each split into 3 levels. Such an approach is known to be inefficient and can lead to optimistic performance estimates[95].

Methods for measurement of performance varied, though no paper provides an assessment of model calibration, an important component of performance for probabilistic models[95]. For discrimination, both DiMasi *et al.* and Lo *et al.* report the commonly used area under the receiver operator curve (AUC). Heinemann *et al.* report the F-measure as their primary performance measure for internal validation and provide no measure of discrimination. For validation on an independently collected dataset, they provide a graphical assessment of discrimination *via* correspondence between the rank of the model predictions and the frequency of approvals. Because a different measure is used for internal and external validation, they cannot be compared. DiMasi *et al.* also report sensitivity and specificity of their algorithm at various scoring cut-offs.

Performance evaluation is again varied and suboptimal (Table 4.2). Though DiMasi *et al.* incorporate bootstrapping and cross-validation into some of the machine learning methods used to select predictors for inclusion in the final model, only the apparent performance (i.e. the performance on the data on which it is developed) of the final model is provided. Lo *et al.* use split-sample validation (70% training, 30% testing) as the primary means to assess performance as well as conducting a preliminary evaluation of the model on independent data for which the outcomes are not yet known, the results of which are presented graphically. A comparison in performance to a slight variation (due to uncertainty in predictor variable definition) of the DiMasi model is also provided. Heinemann *et al.* evaluate performance using 10-fold cross validation as well as externally validating the model in independent data for which the outcomes were recently released.

Table 4.2: Modelling approaches: development, performance, and evaluation of included papers

Item	**DiMasi *et al.*, 2015**	**Lo *et al.*, 2017 (P2APP)***	**Lo *et al.*, 2017 (P3APP)***	**Heinemann *et al.*, 2016**
Type of study	Development	Development, internal validation, external validation of DiMasi model	Development and internal validation	Development, internal validation and external validation
Development				
Handling of missing data	Complete-case analysis	Evaluation of several imputation methods, ultimately using 5-NN approach		N/A
Description of modelling method	Random forest with recursive feature elimination, classification and regression tree analysis (to optimise cut-offs for continuous variables), tests for statistical significance of univariate associations and some multivariable logistic regression. Results used to select important predictors and optimal cut-points for continuous variables in a scoring algorithm	Several models evaluated and optimised for best performance using 10-fold cross-validation: penalised logistic regression, random forests, support vector machine with radial basis functions, and decision trees C5.0.		Several models trained and evaluated using 10-fold cross validation: naïve Bayes, decision trees, random forests, support vector machines and binary logistic regression.
Final modelling method	Scoring algorithm based on several analyses	Random forests algorithm		Random forests algorithm

Item	**DiMasi *et al.*, 2015**	**Lo *et al.*, 2017 (P2APP)***	**Lo *et al.*, 2017 (P3APP)***	**Heinemann *et al.*, 2016**
Methods for selecting predictors for inclusion in multivariable modelling	Final algorithm based on predictor importance from other methods. Predictor input into those methods is based on data availability	Near-zero variance predictors removed		Not explicitly stated; all appear to be included
Methods for selecting predictors during multivariable modelling	Recursive feature elimination in the machine learning methods but not in the final model which is defined by cut-points for the pre-specified variables	Full model approach assumed		Full model approach assumed
Treatment of continuous variables	Categorised in final model	Standardised prior to experiments and assumed to be kept as continuous		Assumed to be kept as continuous
Performance				
Calibration measures	Not provided	Not provided		Not provided
Discrimination measures	AUC	AUC		In external test set, graphical comparison of rank vs. frequency of approval
Classification measures	Sensitivity and specificity at various cut-offs	Not provided		F-measure vs. baseline F-measure given a-priori distribution of data

Item	**DiMasi *et al.*, 2015**	**Lo *et al.*, 2017 (P2APP)***	**Lo *et al.*, 2017 (P3APP)***	**Heinemann *et al.*, 2016**
Evaluation				
Methods for testing model performance (e.g. cross-validation, bootstrapping, external validation)	Apparent performance is provided for final model	Testing on held-out test data (30%) and preliminary external validation on a dataset without known outcomes[†]		10-fold cross validation plus external validation on dataset collected at a later date
Comparison to other approaches	Not provided	Compared to DiMasi *et al.* 2015		Not provided

** P2APP = phase 2 to approval, P3APP = phase 3 to approval.*

† It is slightly unclear whether a single test on held-out data is used or whether the hold-out method is repeated in what is called random sub-sampling. The paper states that they begin by: "first splitting each into a training set… and a testing set… Subsequently, we train 5NN-RF models for each scenario according to the methodology outlined above. We repeat this experiment 100 times for robustness" and also "We split each dataset ... into two disjoint sets, one training set and one testing set... The testing sets are meant to be out-of-sample datasets to evaluate our models. Therefore, we mask their outcomes (that is, we treat them as unknown) and will access them only at the very end to check our performance." We believe the latter statement clarifies ambiguity in the former and indicates that a single held-out dataset is used per model evaluation.

4.4.5 Reported Results

Different performance measures, methods of evaluation and moments for prediction were used, so comparing performance across models is challenging (Table 4.3). DiMasi *et al.* report an apparent AUC of 0.92 (95% CI [0.81, 1.00]) for their scoring tool which predicts approval from conclusion of phase II to approval. In the only directly comparable result between included papers, the Lo P2APP model reports a lower AUC of 0.78 (95% CI [0.75, 0.81]) from split-sample validation. Although DiMasi model reports a higher AUC than LoP2APP, Lo *et al.* attempt to provide a direct comparison between the models, modifying the DiMasi model slightly due to uncertainty in predictor variable definitions. The analysis shows that the model development approach of Lo *et al.* results in superior performance and that the modified DiMasi model achieves a much lower AUC of 0.69 (95% CI [0.58, 0.80]) in held-out data, indicating high levels of overfitting in the original model. The Lo P3APP model achieves an AUC of 0.81 (95% CI [0.78, 0.83]), again on held out data. A single performance measure cannot be extracted from Heinemann *et al.* and instead performance depends on the time point at which the prediction is made, varying from $F = 0.45 \pm 0.08$ (mean ± standard error of mean) at 10 years ahead of the decision (approval or failure) to $F = 0.67 \pm 0.05$ one year before.

In both the LoP2APP and LoP3APP models, a binary predictor variable: “trial outcome – completed, positive outcome or primary endpoint(s) met” is the most important predictor variable. This is followed by “trial status” (whether the trial was completed or terminated). In the DiMasi model, a variable termed “activity” was by far the most important predictor based on *p*-values from univariate association tests and odds ratios from univariate logistic regression. The definition of activity is slightly ambiguous but appears to include whether the results of a randomised trial were positive or negative and, if the trial was non-randomised, a score based on the tumour response rate. At least for randomised trials, this variable therefore appears to be quite similar to the important predictors in the Lo models. For the Heinemann model, "a combination of relatively high values for normalised publication count, commitment and occurrence of MeSH terms "drug therapy" and "therapeutic use" are reported as having the highest correlation with success, based on an associated rule learning process (not the process used to develop the final model).

Of the three papers, only DiMasi *et al*. present their final model, including listing the required predictors (Table 4.4). Lo *et al.* and Heinemann *et al*. both use the random forests algorithm in their final (or best performing) models, but details of any hyperparameter tuning, or code that allows the reader to make predictions using the model, are not presented; in both cases, the predictors required as inputs for the final model are not clarified.

Table 4.3: Results of included papers

Item	**DiMasi *et al.*, 2015**	**Lo *et al.*, 2017 (P2APP)***	**Lo *et al.*, 2017 (P3APP)***	**Heinemann *et al.*, 2016**
Reported performance (discrimination, calibration, classification)	AUC = 0.92 (95% CI [0.81, 1.00])	AUC = 0.78 (95% CI [0.75, 0.81])	AUC = 0.81 (95% CI [0.78, 0.83])	Depends on time point: The F-measure starts at F = 0.45 ± 0.08 (mean ± standard error of mean) at 10 years ahead of time and increases to F = 0.67 ± 0.05 one year before the decision
Details of relative importance or weight of input variables	Activity > no. of patients in pivotal phase II trial > no. of patient treated worldwide > phase II duration†	“Trial outcome, completed, positive outcome or primary endpoint(s) met”, most important across all, followed by “trial status”		A separate rule learning approach is used to examine which factors are most predictive: a combination of normalised publication count, commitment and occurrence of MeSH terms “drug therapy” and “therapeutic use” has the highest correlation to success
Details of final algorithm	See Table 4.4	Random forests classifier model, though details not provided		Random forests classifier model, though details not provided
List of input variables required for algorithm	Activity, no. of patients in pivotal phase II trial, no. of patients treated worldwide, phase II duration	Unclear what are the final variables included in the model		Predictor classes are: article counts, normalised article counts, authors, research commitment, industry affiliation, MeSH subheadings, normalised Shannon entropy of MeSH qualifiers, biomedical terms count, phase term count. Assumed all included in final model but not stated explicitly

** P2APP = phase 2 to approval, P3APP = phase 3 to approval.*

† as determined by odds ratios from univariate logistic regression for each predictor categorised into two levels (in comparison to three levels in the final model).

Table 4.4: DiMasi *et al.* (2015) scoring algorithm

For each factor, a score is assigned and the scores are summed. Higher total scores are interpreted as a higher probability of approval.

	Score		
Factor	*0*	*1*	*2*
Activity	< 3.0% or negative randomised phase II trial	3.0 - 13.8%	> 13.8% or positive randomised phase II trial
No. of patients in pivotal phase II trial	≤ 37	38-49	≥50
No. of patients treated worldwide	> 302,000	50,000 - 302,000	< 50,000
Phase II duration	> 44 months	21-44 months	< 21 months

4.4.6 Issues with the Heinemann Model

An issue specific to the Heinemann model is worth highlighting in some detail. The moment of prediction for the model is defined by the time prior to the outcome event (approval or failure) occurring. This approach would be valid only if these outcomes occurred at the same time, making the actual moment of prediction the same. However, given the definition of failure (failed in phase II/III clinical trials), failure always happens before approval. Therefore, the moment of prediction will not be the same for the two outcomes which could create artificial differences between the groups. Using this approach to make predictions on data with unknown outcomes would not be possible, because the timing of the outcome would be unknown. A moment that is consistent across the two outcome classes must instead be used. Figure 4.2 illustrates the issue with simulated data.

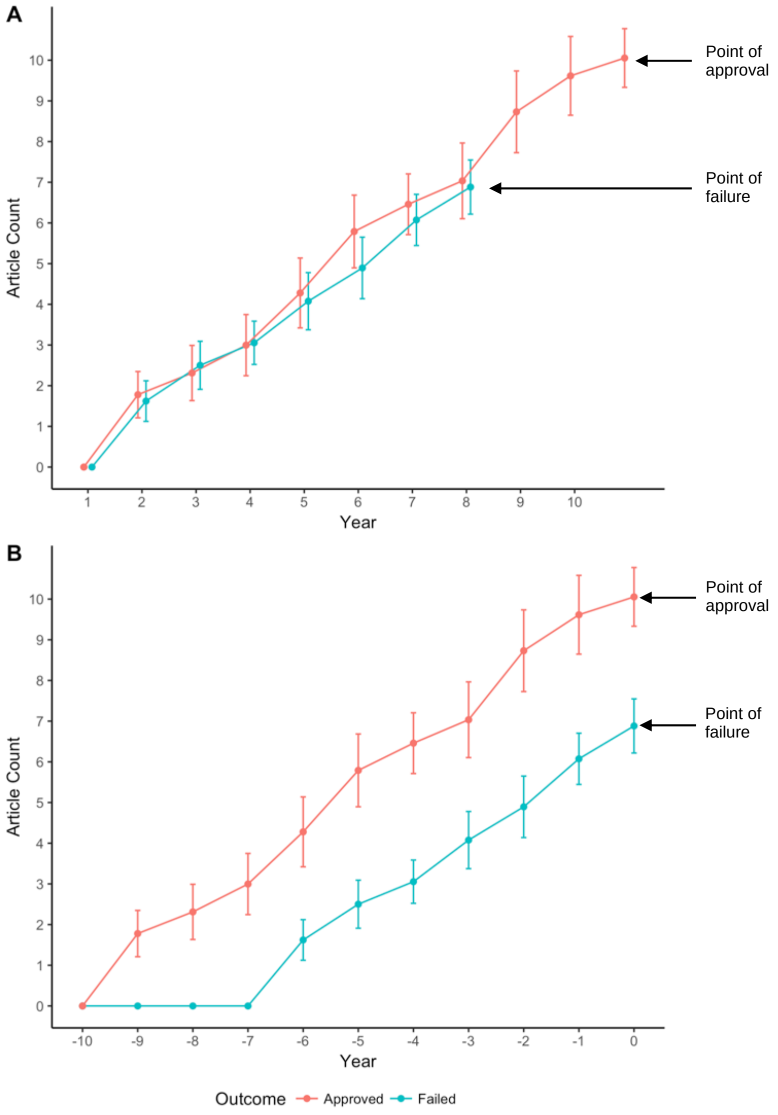


Figure 4.2: Illustration of methodological issues in the Heinemann model

A and B show the exact same data (simulated article count data with the same mean and standard deviation, n = 116 drugs, 42 of which approved as in Heinemann et al.*) but with different definitions of year: A) year from the date of first publication or B) year relative to the outcome (failure or approval), assuming that failure occurs three years earlier than approval. If models are constructed from data treated similarly to B, their ability to discriminate between outcomes could be dramatically over-estimated. Data are mean ± 95% CI.*

4.4.7 Risk of Bias

A number of characteristics of included papers indicate high risk of bias. Risk of bias refers to “the extent that flaws in the design, conduct, and analysis of the primary prediction modelling study lead to biased, often overly optimistic, estimates of predictive performance measures”[122]. Across all papers, reported EPV are lower than generally recommended (Table 4.1), which is likely to lead to overfitting. The maximum plausible EPV across the three papers is 4.5, but this number may be much smaller depending on the number of predictor variables used in modelling, which could not be clearly determined. Also, across all three papers, the methods for testing model performance (Table 4.2) could be improved. Compared to bootstrap approaches for quantifying performance estimates, the apparent performance and estimates from split-sample validation are likely to introduce greater bias[129,130]. In very small datasets, such as that used by Heinemann *et al.*, bootstrapping is more efficient than cross-validation because it allows model development on all data and allows for direct accounting of overfitting in model development[95]. However, in some instances cross-validation can be the preferred approach[129].

The datasets are of varying quality and representativeness. Lo *et al.* utilise by far the largest dataset, representing a very large proportion of approved and failed drugs from the time point and time period of interest. Because they estimate missing data rather than including only complete cases in analysis, the dataset is unlikely to introduce bias into the model and estimated performance. DiMasi *et al.*, on the other hand, include only compounds from the pipeline of top 50 biopharmaceutical companies. These compounds may not be representative of compounds from other companies or settings, so the generalisability of the model is limited. Additionally, only complete cases are included in analyses, resulting in a smaller sample size being used for many analyses and a high risk of biased performance estimates and model specification. Heinemann *et al.* do not appear to have any missing data, though the methods for selecting drugs for the dataset are not detailed and the representativeness of the sample is therefore hard to evaluate. Both the Heinemann and DiMasi datasets are very small (n = 116 and 98 drugs, respectively), which is likely to result in overfitting[95].

In the Heinemann *et al.* model, the use of the timing of the outcome in determining predictor values leads to a very high risk of bias and is likely to lead to overestimation of performance (Figure 4.2). In addition to the issues above, the use of predictor selection strategies in combination with very low EPV and categorisation of continuous variables by DiMasi *et al.* are likely to introduce bias into performance assessment and estimates of predictor-outcome associations, and bias in their performance estimate was demonstrated by Lo *et al.* The Lo *et al.* models exhibit the lowest risk of bias.

4.5 Discussion

A systematic review of publications of multivariable algorithms for the prediction of drug approval identified three relevant papers[125–127] of varying methodological quality and reported performance. Here, we summarise the findings in terms of reporting quality, performance, predictors used in the analysis, other relevant literature, and limitations. Some relevant patent literature is also briefly discussed.

4.5.1 Reporting Quality

Many essential methodological items or details of the final models are not provided. Because none of the three papers provide the full datasets used to develop the models, the only feasible way for readers to use or validate the models is *via* their reporting directly in the papers (for example, reporting of regression coefficients for a regression approach). However, only DiMasi *et al.* present their final model in a format which might allow others to use it (Table 4.4), though some of the predictors are not clearly defined. The Lo and Heinemann models are random forests algorithms, which are not straightforward to present since they are “black box approaches”; however, computer code could be provided that allows the reader to make predictions on their own data or provided in standardised formats such as the Predictive Modelling Markup Language (PMML, http://dmg.org/). Aside from the calculations needed for the model, explicit lists and descriptions of how to collect input variables required for the final models are not provided (Table 4.3), making even model re-development challenging or impossible. None of the papers reported an assessment of model calibration: a measure of the agreement between predicted and observed outcomes and an important component of performance evaluation.

Reporting quality for development of the datasets used to derive the models is varied. Lo *et al.* describe in detail the approach taken to arrive at their final datasets and provide flow charts illustrating data flow. The criteria and approach for the DiMasi and Heinemann datasets are not clear enough to be replicated.

4.5.2 Predictors of Approval

In both the Lo and DiMasi models, the most important predictor variable is similar and essentially reflects whether the phase II trial has a positive outcome. The primary determinant in advancing from phase II to phase III trials is surely the result of the phase II trial. Indeed, the odds ratio (OR) for approval for “activity: high vs. low” (the measure of positive outcome[a]) in the DiMasi dataset is 65.75, which is nearly 20 times higher than the OR for any other presented variable. The AUC provides a measure of discrimination between classes over what would be achieved by random guessing, though random guessing may not be a reasonable comparator in this case. Positive results in the phase II trial will often effectively be a filter for entry into phase III for many drugs, so by including all phase II trials rather than only those which are selected to start phase III, performance estimates might be overestimated in comparison to the situation without use of the model. Even if the presented AUC are accurate, they may therefore not reflect the actual utility of the model over the *status quo*. Furthermore, these predictors cannot be used for drugs earlier in development than the completion of phase II. Since attrition during phase I and II is also high[111], predictors of approval that could be used prior to entry into clinical trials or earlier in clinical development might be useful.

4.5.3 Comments on Performance

Methodological issues in both the Heinemann and DiMasi models render performance estimates likely to be biased (Section 4.4.7), and the reported performance should be interpreted with caution. The approach taken by Lo *et al.* is more appropriate, though some caveats are worth noting. EPV are low, so the model may be overfit. Although the degree of overfitting should be captured to some extent by the internal validation procedure, random variation in the sample used for model evaluation can introduce bias

---

[a] In the case of randomised studies at least, if the primary endpoint was met trial activity was considered "high".

into model performance estimates[129]. Also, as stated above, the high reported AUC may not reflect the utility of the model. Since calibration is not assessed in any of the three papers, the degree of agreement between predicted probabilities and the probability of the outcome occurring is not known. The models may therefore be more useful in ranking and/or discriminating between drugs destined to fail or to be approved than in directly estimating outcome probabilities.

4.5.4 Other Drug Success Literature

Other attempts to predict or describe drug success which do not meet the inclusion criteria for this systematic review are worth mentioning. Several papers aim to relate characteristics of responses identified in clinical trials for oncology drugs to regulatory approval *via* univariate analysis[131–134]. They have small (< 100 drugs) sample sizes and the predictor variables overlap with those used in the Lo and DiMasi models, and are therefore not discussed further. Both Lopes *et al.*[135] and Schachter[136] present relevant models which were excluded during our full text review.

Lopes *et al.* develop a model to distinguish between compounds that are safe and those that are harmful based on characteristics of their protein targets. Though they do not explicitly aim to predict approval, the model is created based on analysis of targets of approved in comparison to problematic (discontinued in development or withdrawn from the market) drugs and is applied at one point to the problem of distinguishing approved and problematic drugs. An AUC of “close to 0.7” is reported. A patent application has been submitted related to this work[137]. Schachter develops a Bayesian belief network model to calculate the probability that a new chemical entity (NCE) will be approved or will fail based on aggregated data on prior success rates and data from early studies of the NCE in question. It is not developed based on a dataset of approved and discontinued drugs, and no estimates of performance are presented, but it does represent one of the earliest attempts to predict the outcome of specific drug candidates rather than simply to describe success rates.

Finally, a large body of literature focussing on physicochemical properties of low molecular weight compounds which aim to assist in prioritisation and improve attrition exists[117,118]. They can broadly be divided into physics-based and empirical methods. Empirical methods are particularly relevant because they are based on identification of quantitative patterns in historical data, similarly to the models discussed in this paper. Although there are recognised issues with some of these approaches[138], it is generally agreed that they have some value and at the very least are widely used[117]. Perhaps because all of the models reviewed herein include both large and small molecules, and therefore comparable information would not be available across molecules, no information on the physicochemical parameters of the drugs were incorporated into the models. Future efforts combining these two disparate approaches might be worth exploration.

4.5.5 Patent Literature

Though we did not conduct a systematic search of the patent literature, key word searching identified one patent application[139] and one granted patent[140] that may be relevant. The application "refers to a method for evaluating, comparing and selecting entities over a broad variety of technical fields. Preferably, the entities are pharmaceutical drugs." A graphical method for selecting or ranking candidate compounds appears to be used, based on comparison of some of their characteristics to approved drugs. The application states that the invention may be used to address "predicting success and failure of a compound". The granted patent's abstract includes the statement: "Methods and systems for determining the selection criteria that in its embodiments can distinguish compounds that successfully meet an objective from those that do not". It does not explicitly give approval as an example of this objective but does provide examples of attributes that might often be required for approval, such as oral bioavailability. The examples provided focus on physicochemical parameters as predictors. These two documents are provided as examples to highlight that commercial applications of models for predicting drug success might benefit from more detailed patent literature analysis.

4.5.6 Limitations

One[127] of the three papers included in this analysis, Lo *et al.*, was not identified *via* any of our database searches and was instead found by manual internet searching. This is not because the search terms were not sufficient to capture the article (this was checked by searching the article title in all databases), but because the article is hosted on a pre-print server (www.ssrn.com) which is not indexed in any of the databases. Note that pre-print publications are a distinct concept from "articles ahead of print": articles made available online by the publisher ahead of journal publication.

Searching pre-print databases, to our knowledge, is not generally recommended during systematic literature reviews. Given the increasing prevalence of pre-print article publication, however, some systematic reviews may wish to consider this. Some attempts have been made to develop tools for easier searching of pre-print literature[141], and searches of databases such as Google Scholar can identify pre-print articles.

Whether or not future reviews include systematic searching of pre-print databases, the identification of a relevant additional article in this chapter highlights the importance of conducting supplementary searching.

4.5.7 Conclusion

We identified very few multivariable models aiming to predict drug approval and the utility of those identified is limited. Methodological issues render the likelihood of bias very high in two models[125,126], and the lack of presentation of a usable model or its predictors render validation or use of the third challenging[127]. Across all models there is considerable room for improvement in methodological and reporting quality. Despite this, promising performance measures are reported and, in the case of Lo *et al.*, the internal validation procedures used should limit over-optimism in their assessment. With improvements in modelling procedures, and perhaps with a combination of predictors used here as well as others associated with drug success such as physicochemical properties, it may be possible to develop and report usable models that improve attrition.

5 Development and Internal Validation of Multivariable Prediction Models for Small Molecule Drug Approval

5.1 Abstract

Several efforts to predict regulatory approval have been made, though no multivariable models to date include physicochemical predictors or attempt to predict approval from the preclinical stage. Therefore, we developed and internally validated two multivariable logistic regression models for predicting regulatory approval of small molecule therapeutics. The first (n = 1,220 drugs, 212 approved) incorporates physicochemical parameters, commercial factors, literature counts, and economic variables as predictors. It achieved an optimism corrected performance of 0.59 area under the receiver operating characteristics curve (AUC; 95% CI [0.55, 0.63]). The second model includes only physicochemical parameters (n = 3,229 drugs, 904 approved) and achieved an AUC of 0.62 (95% CI [0.60, 0.64]). Sampling bias is introduced during dataset development, so we evaluate the impact of changes in outcome distributions and show that the models are relatively robust to those changes. Further work is required to externally validate the model and evaluate its impact, financially and to patients.

5.2 Introduction

Developing drugs is exceptionally expensive, with estimates as high as \$2.6 billion per new chemical entity[3]. One of the key drivers of this cost is the high failure rate observed during clinical trials[142], in which candidate drugs are tested to determine whether they are safe and efficacious. As well as a high financial cost, there is a direct cost to human health, as patients are often exposed to interventions that are ultimately deemed unsuitable and which may be harmful. The ability to prioritise drug candidates that are eventually approved, compared to those that are considered for clinical trials but eventually fail, could therefore be of considerable value.

Several attempts have been made to develop predictive models of regulatory approval or clinical development on the basis of a range of predictors[125–127]. However, these models have not included detailed information related to the physicochemical parameters of the compounds, which could be informative and which have been the basis of many historical attempts to distinguish potential drugs from other chemicals[143–145]. Here, we attempt to consolidate these two approaches and develop a dataset comprising both physicochemical parameters and other parameters that could influence the success of a compound, such as information about the company developing it, literature on the drug and the general economic environment. To allow comparable physicochemical data to be collected, we limit the scope of the predictive model to small molecules. To maximise the potential value of the model, we attempt to predict regulatory approval from the preclinical stage.

Many attempts to distinguish drugs from non-drugs on the basis of physicochemical parameters have focussed only on comparing approved drugs to background chemicals (i.e. reagents thought to have little therapeutic relevance)[118,145], rather than to discontinued drugs. However, most of the cost of drug development is due to attrition, so the ability to distinguish compounds that have entered the drug development process could be more valuable, though is likely more challenging (Figure 5.1). Using untested compounds as a comparator is suboptimal because we do not know if the compounds would be drugs if tested in trials. Indeed, analysis has shown that up to 40% of compounds in “non-drug” databases have chemical similarity to approved drugs[146]. Other methods have compared approved drugs to experimental drugs (i.e. drugs that are at the preclinical or animal testing stage)[147] but these are again less than ideal because some of the experimental drugs are likely to eventually be approved. Therefore, we also develop a larger dataset comprising only physicochemical information relating to approved and discontinued drugs and develop a model to distinguish between them.

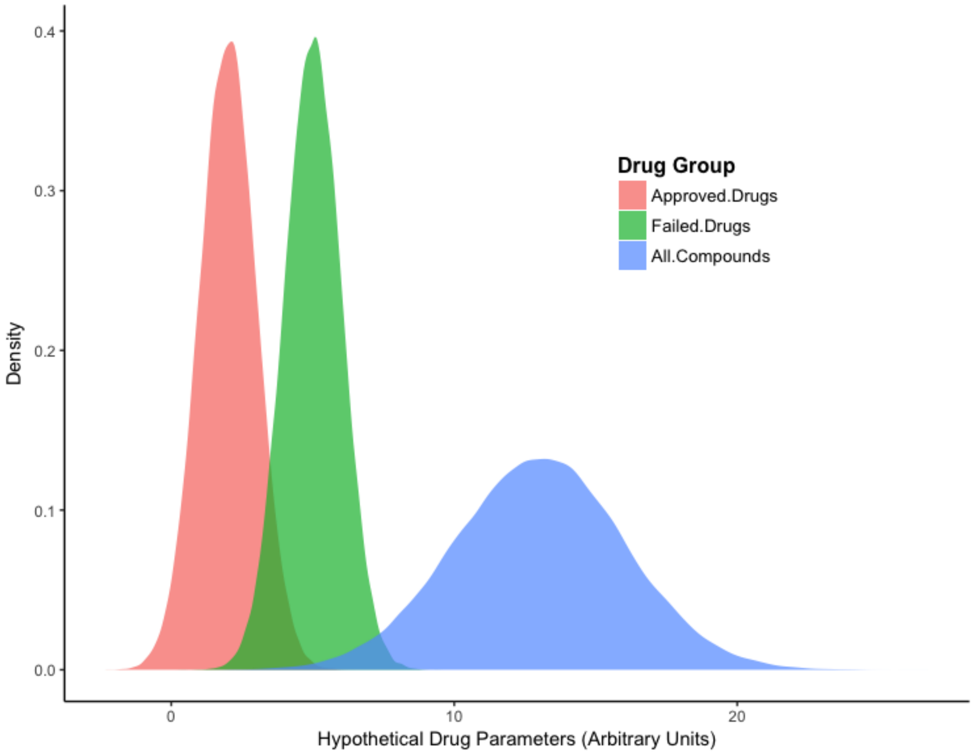


Figure 5.1: Comparing approved and failed drugs to all compounds

Illustration of the typical similarity of approved drugs, failed drugs (i.e. drugs that entered the clinic and were not eventually approved), and all compounds (i.e. those that may or may not have been considered for use in humans) for hypothetical drug parameters. Approved and failed drugs are generally more similar to each other than to all compounds, but many existing computational attempts to predict drug success focus on distinguishing approved drugs from general background compounds.

In this chapter, we develop two distinct models, both of which may have utility in improving attrition rates: i) to distinguish between approved and discontinued drugs based on information available prior to the commencement of clinical trials (referred to as the “discovery model” based on the “discovery dataset” of 1,220 drugs) and ii) to distinguish between approved and discontinued drugs based solely on physicochemical parameters (referred to as the “physicochemical model” based on the “physicochemical dataset” of 3,229 drugs). In both cases, logistic regression models are fit and internally validated. Following correction for optimism in performance assessment, we find the second model performs slightly better. The model also performs much better than a commonly used drug-likeness “rule” in our dataset. We provide a more detailed comparison to other relevant literature in the discussion.

Throughout this chapter, we aim to follow the relevant reporting and methodological recommendations of the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement[124] and accompanying explanation and elaboration[95]. Though TRIPOD is designed for clinical prognostic or diagnostic studies, many of the details are relevant.

5.3 Methods: Discovery Model

Here, we attempt to develop a prediction model for small molecule regulatory approval based on predictors available prior to clinical trials of those compounds. The development of the dataset is described, followed by the modelling procedures. R v3.4.4 (ref: [148]) was used for any analysis conducted in R. The “tidyverse” packages, v1.2.1 (ref: [149]) were used throughout for analysis in R.

5.3.1 Included Compounds

Compounds from Cortellis™ Competitive Intelligence (Cortellis) were included in the dataset used to develop the model[b]. Cortellis is a commercial database containing drug development and clinical trial information on more than 70,000 drug programs (https://clarivate.com/products/cortellis/cortellis-competitive-intelligence/). Only drugs included in this database were used because we wanted to distinguish between approved drugs and compounds which were of therapeutic interest but which eventually failed, rather than between approved drugs and other compounds which may or may not be considered promising drug candidates. Only compounds known to be under investigation as therapeutics are included in this database, thus satisfying this requirement.

To be included, a compound in the Cortellis dataset had to meet the following criteria:

1. Be described in the dataset as a small molecule therapeutic, and not be a combination drug, peptide, peptidomimetic, or polysaccharide
2. Be either approved (denoted as "launched" in Cortellis) for at least one indication in any jurisdiction or discontinued across all indications and jurisdictions (see section 5.3.3 Outcome Variable). Discontinuation refers to drugs whose development is stopped prior to market approval, and does not include drugs withdrawn from the market

---

[b] Access to the database kindly provided by Dr David Brindley.

3. Have a “discovery” date in the database which was the earliest date listed for the drug, referred to here as the start date. The start date in this case approximates the date that a compound was first publicly announced as being under development and precedes entry into clinical trials. This date is required for any time dependent variables.

We aimed to include all drugs that met these criteria and did not put any restrictions on sample size.

5.3.2 Moment of Prediction

To maximise the potential value of the model, we wanted to make predictions from the earliest time point possible for drugs under investigation as therapeutics. However, we also needed to be able to gather data from the public domain. Therefore, the date on which a drug is first publicly announced as under investigation was used as the moment of prediction, as long as that date preceded entry of the drug into clinical trials. This is indicated in Cortellis by the “discovery” date (here the “start date”). To ensure that this time point was consistent, we did not include candidates which had entered clinical trials at the time they were first publicly announced. The moment of prediction for this model, therefore, roughly corresponds to preclinical development.

5.3.3 Outcome Variable

The outcome we aimed to predict was approval for at least one indication in at least one jurisdiction or drug discontinuation across all indications and jurisdictions. This was determined by the development status listed in Cortellis.

5.3.4 Data Cleaning

Drug names, as listed in the dataset, often included information relating to the indication being investigated, the formulation, or the company developing the drug, and were therefore cleaned to remove this information. This resulted in the presence of a number of duplicate drug names, of which those with the earliest date were retained. Combination drugs, peptides, peptidomimetics, and polysaccharides were removed. These steps were automated in R. After collection of physicochemical data (Section 5.3.5.1), compounds with molecular weight > 1,500 Da were manually reviewed and if appropriate removed from the dataset. This was sometimes the case when they had not been labelled appropriately in the original Cortellis data.

5.3.5 Candidate Predictor Variables

Candidate predictor variables were selected on the basis of their inclusion in previous, similar studies, logical reasoning, and the plausibility of collecting information on them given the data we could access. They fell into the following five categories:

1. Physicochemical parameters (eight candidates)
2. Developer characteristics (three candidates)
3. Drug usage (two candidates)
4. Literature (three candidates)
5. Economic (seven candidates)

The specific predictors are summarised in Table 5.1 and described in detail below along with how they were collected.

5.3.5.1 Physicochemical Parameters

Eight candidate physicochemical parameters were selected (Table 5.1) as used in the model developed by Bickerton *et al.*[145], a seminal paper attempting to quantify drug-likeness or "chemical beauty" based on analysis of approved drugs. Several of these parameters have long been considered to be associated with drug-likeness[143,150].

Physicochemical information was collected using R and KNIME (v3.5.3). The R package *rpubchem* (v1.5.16) was used to search and download available physicochemical information from PubChem, a public repository containing information on millions of compounds[151]. Drug names from Cortellis were used to search PubChem to retrieve compound IDs (CIDs), which provide a unique numerical ID for every compound in the PubChem database. When a compound name returned more than one CID, the CIDs were manually reviewed to identify the correct entry. If the correct entry could not be discerned (often they were very similar), one of the returned CIDs was randomly selected. The CIDs were then used to retrieve all available physicochemical information, as well as canonical simplified molecular-input line-entry system (SMILES) structures, which describes chemical structures in a character string. We originally planned to collect, from PubChem: molecular weight, hydrogen bond donor count (HBDC), hydrogen bond acceptor count (HBAC), rotatable bond count (RBC), topological polar surface area (TPSA), and a calculation of the octanol-water partition coefficient (XLogP3 [ref: [152]]). However, XLogP3 had high levels of missing data (21%) and was therefore discarded and calculated directly in KNIME, along with the number of aromatic rings (AROM) for each compound. Canonical SMILES downloaded from PubChem were used to generate a molecular format

(CDKCell), from which the RDKit descriptor calculator calculated AROM and the XLogP node calculated XLogP[153]. Though we had intended to also calculate the number of structural alerts for each compound as per Bickerton *et al.*[145], the method was not clearly described and we were unable to do so.

5.3.5.2 Developer Characteristics

Along with the drug name and development status, developer characteristics and information on drug usage were retrieved from Cortellis. Developer characteristics comprised the name of the originator organisation (for example the company or university from which the drug originated) and the organisation(s) developing the drug. These were used to generate three predictor variables: a binary variable indicating whether the drug originated from a commercial or academic organisation, a binary variable indicating whether the organisation developing the drug was different from the originator organisation, and a numerical variable representing the number of distinct organisations developing the drug. We did not account for organisational name changes (e.g. due to mergers) in generating these variables, which was automated using R.

Table 5.1: Description and rationale for candidate predictor variables for discovery and physicochemical models

Predictor Variable	**Abbreviation**	**Rationale**	**Variable Type**	**Model***
Physicochemical Parameters				
Molecular weight, g/mol		Higher molecular weight leads to lower bi-layer permeability. Component of Lipinksi's Ro5	Numeric	D, PC
Hydrogen bond donor count	HBDC	High number thought to reduce membrane bi-layer permeability. Component of Lipinksi's Ro5	Numeric	D, PC
Hydrogen bond acceptor count	HBAC	High number thought to reduce membrane bi-layer permeability. Component of Lipinksi's Ro5	Numeric	D, PC
Logarithm of the octanol/water partition coefficient (XLogP method)	XLogP	Measure of lipophilicity and component of Lipinski's Ro5	Numeric	D, PC
Rotatable bond count	RBC	Measure of molecular flexibility, additional rule later added to Ro5	Numeric	D, PC
Topological polar surface area, Å^2	TPSA	Lower values thought to improve adsorption and permeability. Included in efforts to quantify drug-likeness	Numeric	D, PC
Number of aromatic rings	AROM	Smaller number of aromatic rings associated with drug success. Included in efforts to quantify drug-likeness	Numeric	D, PC
Number of structural alerts		Indicate structural components that may be mutagenic, reactive, or have unfavourable pharmacokinetic properties. Included in efforts to quantify drug-likeness	Numeric	None

Predictor Variable	**Abbreviation**	**Rationale**	**Variable Type**	**Model***
Developer Characteristics				
Academic vs. commercial origin	Academic commercial	Plausible that there could be differences in likelihood of success in academic vs. commercial origin	Binary	D
Non-originator company developing drug	Company change	Indicates licensing or acquisition, which may represent external validation of concept and therefore increased likelihood of success	Binary	D
Number of organisations developing drug	Distinct companies	May provide indicator of commercial interest in drug	Numeric	D
Drug Usage				
Route of administration	ROA	Different ROA drugs may have different parameters for success associated with them. Included as predictor in other attempts to predict approval	Categorical, three levels†	None
Indication		Approval rates differ across indications. Included as predictor in other attempts to predict approval	Categorical, 14 levels‡	None
Literature				
Number of submitted patents	Submitted patents	Potential indicator of commercial interest and have been associated with predictions of drug status previously	Numeric	D
Number of granted patents	Granted patents	See “submitted patents”	Numeric	None
Number of citations (academic literature articles)	Citations	Indicator of scientific activity which has been associated with drug status	Numeric	D

Predictor Variable	**Abbreviation**	**Rationale**	**Variable Type**	**Model***
Economic[§]				
Pharma R&D spend (100M USD, adjusted to 2015)	RandD	Pharmaceutical companies are a major source of funding for drug development, and the extent of their spending on R&D could influence approval	Numeric	D
US interest rate	US interest	The US is a major market for drug discovery and development, and interest rates may affect the availability and cost of capital	Numeric	D
UK interest rate	UK interest	The UK is a major economy and location for drug development in the EU. As above, interest rates may affect the availability and cost of capital	Numeric	None
Number of companies receiving venture capital in EU (10s of companies)	EU VC companies	The availability of venture capital could impact early-stage therapeutics, particularly outside of big pharma	Numeric	D
Venture capital spending in EU (M EUR, adjusted to 2015)	EU VC M EUR	See EU VC companies	Numeric	None
Euro Area gross domestic product change (annual %)	EA GDP	Measure of general economic health in major market	Numeric	D
US gross domestic product change (annual %)	US GDP	See EA GDP	Numeric	D

** Model in which the candidate predictors were eventually included. D = discovery, PC = physicochemical.*

† Levels were: oral, no ROA provided, other ROA

‡ Levels were: infectious and parasitic, neoplasm, blood and blood-forming organs and disorders involving the immune system, endocrine and nutritional and metabolic diseases, mental and behavioural, nervous system, eye and adnexa, circulatory system, respiratory system, digestive system, skin and subcutaneous tissue, musculoskeletal system and connective tissue, genitourinary, other.

§ All economic predictor variables are annual figures for the year prior to the start date of the drug. This is because data for the full start date year would not be available at the time of prediction.

5.3.5.3 Drug Usage

Two candidate predictor variables were collected relating to drug usage: indication(s) and route of administration (ROA), both of which have been used as predictor variables in other models[127]. Indications were manually assigned to 14 groups derived from the World Health Organization International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10; ref: [154]). Drugs could be assigned more than one group if being investigated for multiple indications, resulting in 14 binary predictor variables. For ROA, each drug was assigned to at least one of the following categories: inhaled, oral, parenteral, topical, or no ROA provided. Because the number of inhaled and topical drugs was very small (< 2% of drugs in each), these were combined with parenteral to create an “other ROA” category, resulting in three binary predictor variables: oral, other ROA, or no ROA. Generation of both indication and ROA variables was automated in R.

5.3.5.4 Literature Information

The published literature has been used in attempts to predict the success of drugs for many years[155] and there is some evidence that changes in patent or academic literature are associated with drug success[90,156]. Therefore, we aimed to collect three candidate predictors: the number of academic articles, the number of submitted patents, and the number of granted patents. In all cases, we were interested only in the numbers available prior to the date at which the prediction was being made. A script was written in Python (v3.6) which automatically downloaded CSV files from PubChem for each

CID and counted the number of articles or patents available prior to the date of interest[c]. For submitted patents, the date in PubChem corresponds to the date that the patent was submitted. However, a patent is generally not published for about 18 months after submission[157], and we therefore only recorded submitted patents available 18 months prior to the date from Cortellis. Where no data were provided in PubChem, it was assumed that the count was zero.

5.3.5.5 Economic Variables

We postulated that variables broadly describing the economic and investment climate at the time a drug candidate is initially being developed could affect the likelihood of success. We therefore collected candidate predictors to reflect the availability of venture capital for early-stage ventures, pharmaceutical research and development (R&D) spending, and the general economic environment. Data for the year prior to the start date were used for each drug because data for the full year in which a prediction was being made would not be available at the time of prediction.

5.3.5.5.1 Venture Capital

Data on the number of investments and amount of early-stage venture capital spending were generated for 14 European countries (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Netherlands, Portugal, Spain, Sweden, and

---

[c] This script was written by Piotr Pawlik, a student volunteer at the SENS Research Foundation (Mountain View, CA).

United Kingdom). Note that this included investments in high tech industry and knowledge-intensive services as well as biotechnology and pharmaceuticals.

Data on EU venture capital spending and number of investments were downloaded from Eurostat[158] and were available from 1989 to 2015. Prior to 2007, data were indexed differently, so to generate comparable data, investments in seed and start-up stage companies for 2007 to 2015 were combined to generate a total sum of early-stage investment capital deployed. From 1996 to 2015, Harmonised Index of Consumer Prices (HICP) data from Eurostat[159] were used to convert investments for each year and in each country to 2015 euros to account for inflation. From 1989 to 1996, these data were not available so inflation rates[160] for each year were used to estimate the HICP.

The number of early-stage companies receiving investments was also extracted. As with spending, data on seed and start-up stage companies for 2007 to 2015 were combined to generate a total number of companies. From 1989 to 1995, the number of companies invested in was not available but the number of investments made was. To estimate the number of companies in which investments were made, data from 1995 to 2006, in which both the number of companies invested in and number of investments made were available, was used to generate a linear regression model (Figure 5.2, β = 0.73, 95% CI [0.62, 0.83], $R^2 = 0.97$). The model was used to estimate the number of companies invested in from 1989 to 1995 from the number of investments made in that period. For both variables, no data were available prior to 1989, so values were imputed by the mean from 1989 to 1997 data, during which there was little change (data not shown; mean imputation applied to only 1% of included drugs [n = 12]).

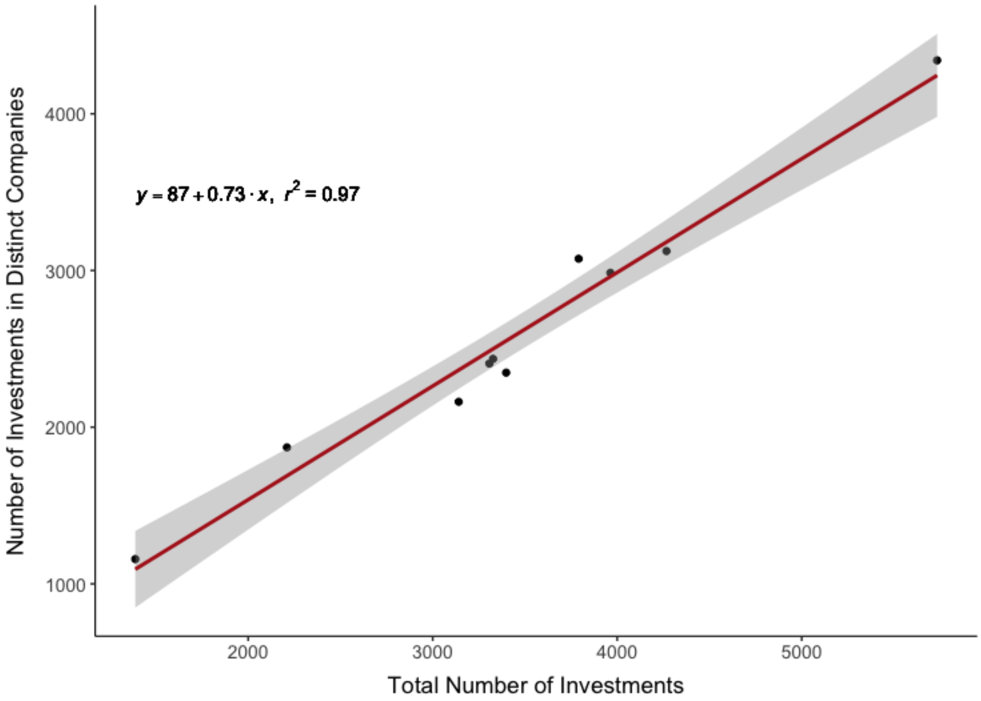


Figure 5.2: Investments vs. companies, 1996 – 2006

Regression model of the number of investments made in 14 European countries from 1996 – 2006. Red line is the linear regression line (± 95% CI). The model was used to estimate the number of companies invested in from 1989 to 1995 from the number of investments made in that period.

5.3.5.5.2 Pharmaceutical R&D Spending

Data were collected on pharmaceutical R&D spending from the annual Pharmaceutical Research and Manufacturers of America (PhRMA) membership survey[161] and adjusted to 2015 USD using consumer price index (CPI) data. PhRMA data was available to 1970, and one drug in our dataset had a date of 1967: 1970 data was used for this drug. Ideally, global pharmaceutical R&D spending would have been included in the model but this information was not available across the required time period. For 2008 to 2015, however, independent estimates of global R&D spending were available[162], and these were highly correlated with the PhRMA results (Pearson product-moment correlation, $r = 0.87$, $p = 0.005$). It is therefore unlikely that including global data would have added significant additional information.

5.3.5.5.3 Macroeconomic Variables

Interest rates provide an indication of the cost of money and were collected for two major economies (US and UK). For the US, the effective federal funds rate was used[163], which is often considered to be the most influential US interest rate, and for the UK the Official Bank Rate was used[164]. In both cases, yearly data were generated by taking averages over the time intervals available. Annual gross domestic product (GDP) change data for the Euro area and the US were downloaded from the World Bank website (data.worldbank.org/indicator/NY.GDP.PCAP.KD.ZG) as they could provide a measure of the overall economic health of those areas.

5.3.6 Variable Selection

A full model approach with no variable selection during or prior to model development was planned. However, after data collection there were too few events per variable (EPV; where events are the number of outcomes in the smallest category, and variables are the predictor variables) to retain all predictors. For a binary outcome variable, it is generally recommended that a minimum of 10 EPV are used[165,166], and more EPV further reduce the likelihood of overfitting[167]. Therefore, we reduced the number of variables based on collinearity[168], the likely relevance of predictors to the outcome[167], and, for categorical variables, the number of categories and feasibility of reducing them. There is a large body of literature indicating that the physicochemical parameters of a drug candidate influence its likelihood of success, so we decided in advance not to remove any physicochemical variables.

5.3.7 Modelling Approach

5.3.7.1 Model Type

We developed and evaluated a commonly used and recommended classifier algorithm, logistic regression[95], with an internal validation procedure. The assumption of linearity in the logit was checked graphically by plotting continuous predictor values against the logit.

5.3.7.2 Performance Measures

Model performance was primarily assessed using the commonly used and recommended c-index[95], also known as the area under the curve receiver operating characteristics curve (AUC, which is the estimated probability that a classifier will rank a positive outcome, i.e. an approved drug, higher than a negative outcome, i.e. a failed drug[127,169]). The AUC is a measure of discrimination and ranges from 0.5 to 1, with 0.5 representing no ability to discriminate between classes, and 1 representing perfect discrimination. The R *ROCR* package (v1.0.7) was used[170]. We also assessed model calibration, which is the agreement between predictions from the model and observed outcomes, by plotting predicted outcome probabilities against observed outcome frequencies[95].

Other attempts to predict regulatory approval from a variety of predictors do so from a later stage in development and are not directly comparable to this approach. Therefore, we do not compare the performance of this model to any other published models.

5.3.7.3 Internal Validation

Bootstrapping was used to estimate the optimism in the predictive performance of the model[171,172] and to calculate confidence intervals for the optimism corrected performance. A model was first developed and fit to the whole dataset, providing the apparent performance. The dataset was then sampled with replacement to generate a bootstrapped sample of the same size as the total dataset. A model was developed on the bootstrapped sample using the same procedure as used for the whole dataset. The performance of this model was quantified on both the bootstrapped sample from which it was developed, and on the whole dataset, and the difference between the performance on the two datasets was then calculated to give the optimism. This procedure was repeated 1,000 times per model, and an average of the optimism calculated and subtracted from the apparent performance, resulting in optimism-corrected performance estimates[95]. Confidence intervals for the optimism and therefore performance were also extracted. Bias in calibration was similarly assessed by bootstrapping with 1,000 resamples using the calibrate function from the *rms* package[173] (v5.1.2) in R and reported graphically.

5.3.7.4 Sensitivity Analyses

5.3.7.4.1 Dealing with Sampling Bias

It is likely that the proportion of approved drugs in our sample differs from the true proportion in the population of preclinical drugs with commercial interest. First, it is likely that, even before data processing, our sample contained a greater proportion of approved drugs than would be present in a truly representative sample, because some drugs are backfilled in the database (e.g. information relating to their development is added after that development has occurred and the drug becomes better known). Second, the process of developing our dataset modified the distribution because we were not able to identify CIDs for every drug name, and approved drugs were more likely to successfully match to a CID, resulting in a greater proportion of approved drugs in our final dataset (0.18) in comparison to that of the drugs meeting the filtering criteria (0.11; see Results, Section 5.5). Most predictive modelling approaches assume that the distribution of training data and the population in which the model will be used is the same, so it is important to address this[174,175].

Ideally, we would have been able to include all drugs from the Cortellis database in our analysis, rather than including only complete cases. When there is missing data (which was the case for all drugs we could not match to a CID), it is usually preferable to impute it to allow the inclusion of those data in the model development process[95]. However, in our data, the commercial information was available for all drugs, and either all or none of the physicochemical information was available for all drugs. It would not be reasonable to attempt to estimate the chemical structure of a drug based purely on commercial information about it, so we did not use this approach. Even if we had been able to impute the full physicochemical data, the sample would still be unlikely to reflect the "true" population of drugs for reasons already discussed.

If the true class distributions are known, under or oversampling can be used to generate samples from the sampled, biased data which reflect the true class distributions and can be used for model development or evaluation[176,177]. In this case, the true distribution of discontinued and approved drugs from the preclinical stage is not known with certainty but is likely to fall below 13.8%: a recent estimate for the success rate of drugs entering clinical trials[8]. Because undersampling of approved drugs during model development would violate the 10 EPV recommendation, we instead undersample the test dataset and evaluate how the predictive performance differs across several thresholds (7.5, 10, 12.5, and 15%) of approved drugs as a proportion of total drugs. This is useful to evaluate the robustness of the model to differences in class distributions. It should be noted that this procedure does not provide an assessment of the optimism in the calculated AUC and associated confidence intervals and is therefore useful only as a comparative measure to the apparent performance. Specifically, the following steps were taken for each level of undersampling:

1. A model was developed using all *n* observations in the original dataset (i.e. the physicochemical model)
2. The approved drugs in the original dataset were undersampled to generate a dataset with the desired proportion of approved drugs
3. The physicochemical model was applied to the undersampled dataset and the AUC was calculated
4. Steps 2 and 3 were repeated 1,000 times
5. The mean of all 1,000 AUCs was calculated and the 95% CI for performance extracted from the results.

5.3.7.4.2 Outliers and Influential Observations

Outliers and influential observations were identified and their impact on the regression coefficients analysed. A Bonferroni outlier test was conducted to identify outliers and Cook’s distance was used to assess influence[178,179]. For highly influential observations, models were constructed with and without the observation and the regression coefficients were compared[179].

5.3.7.4.3 Impact of Non-physicochemical Predictors

To assess the marginal improvement in predictive performance provided by the non-physicochemical predictors, we assessed the performance of a model with just physicochemical parameters compared to a model with all predictors.

5.4 Methods: Physicochemical Model

The filters imposed during development of the discovery dataset resulted in the exclusion of a large number of drugs, so we also developed a larger dataset including only physicochemical predictors, which are available for all drugs that can be matched to a CID. This section describes any differences between the methods used for the discovery and physicochemical model. The same outcome variable, data cleaning (as applicable), variable selection, model type, internal validation method, and outlier and influential observation assessments were used. Refer to Section 5.3 for information on each of these.

5.4.1 Included Compounds

Cortellis was again used to collect drug names and development status. To be included, a compound in the Cortellis dataset had to meet the following criteria:

1. Be described in the dataset as a small molecule therapeutic, and not be a combination drug, peptide, peptidomimetic, or polysaccharide
2. Be either approved (denoted as "launched" in Cortellis) for at least one indication in any jurisdiction or discontinued across all indications and

jurisdictions (see section 5.3.3 Outcome Variable). Discontinuation refers to drugs whose development is stopped prior to market approval, and does not include drugs withdrawn from the market.

5.4.2 Moment of Prediction

The dataset incorporates compounds entering the database at the “discovery” stage (explained in Section 5.3.1) as well as any other stage of development. It is, therefore, applicable to prediction of approval from any moment after a compound is entered into the Cortellis database. In practice, this corresponds to any time-point during or after preclinical development.

5.4.3 Predictor Variables

Only physicochemical parameters were included as predictors, and these were calculated or collected as specified in Section 5.3.5.1 with one difference: when multiple CIDs were identified for a drug name, it was impractical to manually review each of these to determine which CID was correct (1,222 CIDs representing 339 drug names). A review of a random subset and experience with the same problem for the discovery dataset indicated that the physicochemical information for different CIDs for the same drug names were almost always very similar, and it was often not possible to discern which CID was correct. Therefore, we randomly selected a single CID for each drug name and used this for subsequent analysis. Sensitivity analysis to address the impact of the use of different CIDs for these drugs is planned (Outstanding Work, Section 5.6.5).

5.4.4 Modelling Approach

5.4.4.1 Performance Measures

AUC was again used as the primary measure for performance, and calibration was also assessed as described above. We also compared the performance of our model to Lipinski's rule-of-5 (Ro5, ref: [143]), a very commonly used method of assessing drug-likeness based on simple physicochemical data. It should be noted that it was originally developed as a rule for oral bioavailability rather than specifically regulatory approval, though it is commonly used outside of this context during drug discovery for prioritisation of compounds[118]. An updated and commonly used version of the Ro5 states that compounds with at least one violation of the following rules are unlikely to be orally bioavailable[180]:

- Molecular weight ≤500 Daltons
- Oil/water partition coefficient (LogP) ≤5
- Hydrogen bond donor count ≤5
- Hydrogen bond acceptor count ≤10
- Rotatable bond count ≤10

Compliance with these criteria was assessed using KNIME's "Lipinski's Rule-of-Five" node, and the sensitivity and specificity of the rule in the physicochemical dataset then calculated in R.

5.4.4.2 Sensitivity Analyses

5.4.4.2.1 Dealing with Sampling Bias

As with the discovery dataset, the physicochemical dataset likely contains a greater proportion of approved drugs than would be present in an unbiased dataset (See Results, Figure 5.3). Therefore, we perform a sensitivity analysis following the same method detailed in Section 5.3.7.4.1, undersampling the approved class in the test dataset (at 5, 10, 15, and 20% as a percentage of the test dataset) to give an estimate of the performance of the model on more realistic class distributions.

Given that EPV was much higher in this dataset, we were also able to explore the impact of undersampling approved drugs during model development. We conducted an additional analysis by undersampling approved drugs during both training and testing in each bootstrapped sample at the following levels: 5, 10, 15, and 20%. The purpose was to determine optimism corrected performance estimates if the model had been developed on more representative samples. Specifically, the following procedure was used for each level of undersampling:

1. The approved drugs in the dataset were undersampled to generate a dataset of size *s* with the desired proportion of approved drugs
2. A model was developed on the undersampled dataset and its performance evaluated in that same dataset, giving the apparent performance ($performance_{apparent}$)
3. A bootstrapped sample of size *s* was generated by sampling the undersampled dataset with replacement

4. A model was developed on the bootstrapped sample and its performance assessed in that sample, giving the bootstrap performance ($performance_{boot}$)
5. The model generated in step 4 was evaluated in the undersampled dataset generated in step 1, giving the test set performance, $performance_{test}$
6. The optimism in the fit in this sample, o, was calculated as $performance_{boot}$ - $performance_{test}$
7. Steps 2 to 6 were repeated 1,000 times
8. An average of o across all iterations was generated to give the estimated optimism, O, and confidence intervals were also extracted
9. O was subtracted from the average of $performance_{apparent}$ across all iterations to give an optimism corrected estimate of the model performance when developed and internally validated in more representative samples.

5.5 Results

Our data collection procedures resulted in a discovery dataset containing 1,220 drugs, 212 of which were approved, and a physicochemical dataset containing 3,229 drugs, 904 of which were approved (Figure 5.3). Logistic regression models were developed and internally validated on each dataset.

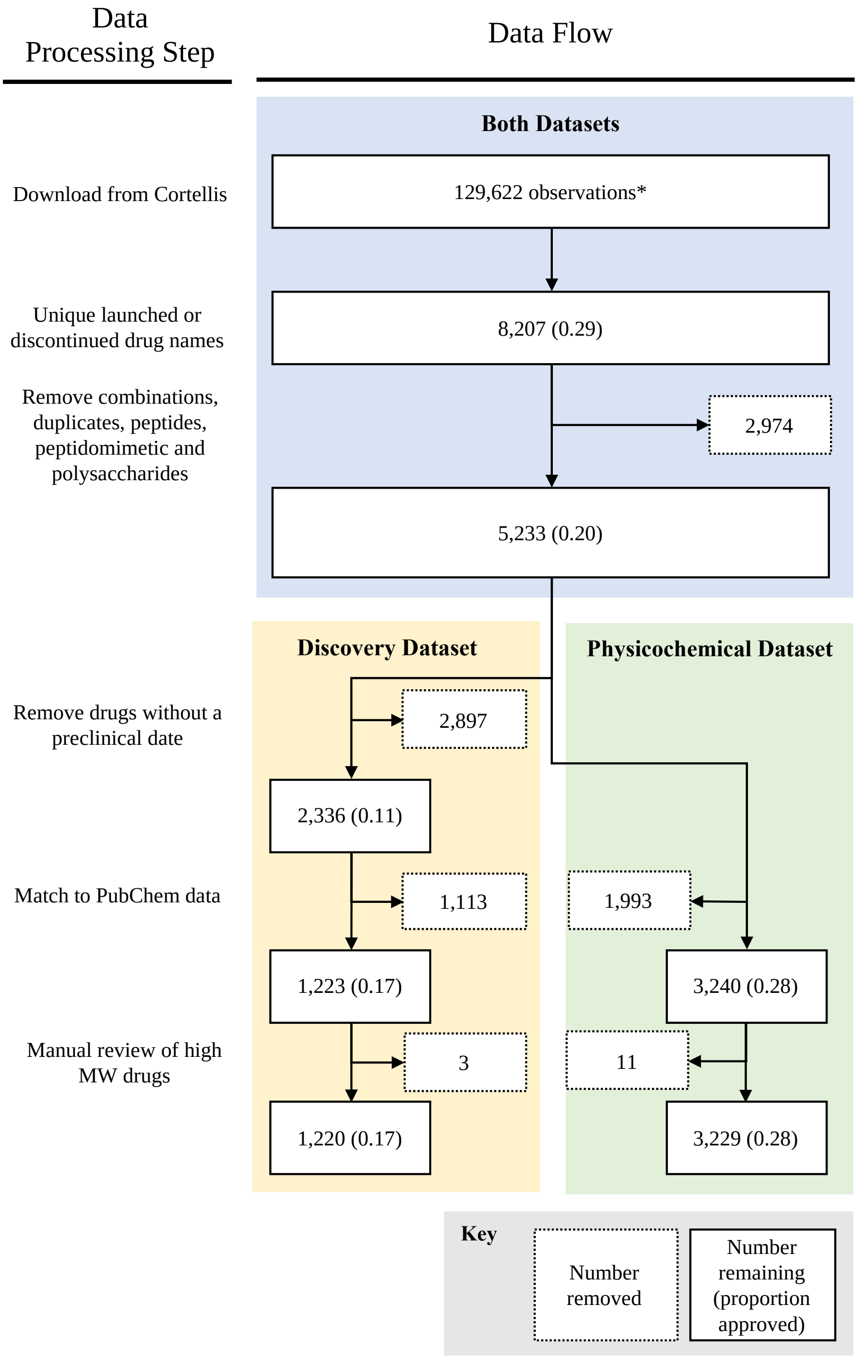


Figure 5.3: Flow of data throughout development of datasets

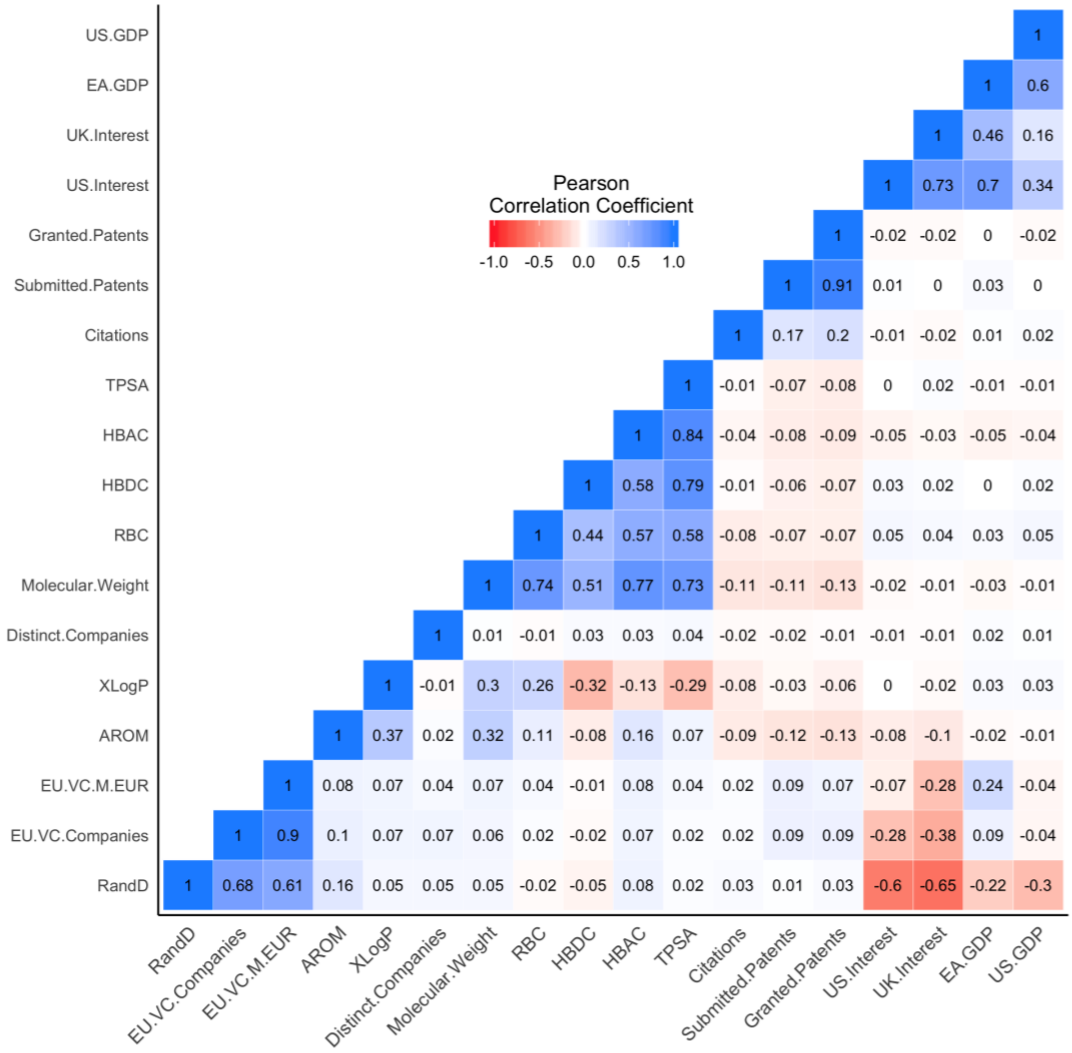
**each drug has multiple rows (observations) associated with it, representing data at a particular point in time, indication, development stage, etc.*

5.5.1 Discovery Model

5.5.1.1 Variable Selection Before Modelling

Originally, a full model approach was planned because there is limited cost to collecting the candidate variables. However, after data collection the total number of drugs was 1,220, of which 212 were approved, so the number of variables was reduced to increase EPV from 5.8 to at least 10.

Indication group had a high number of categories (14) which could not be combined to a generate a manageable number meaningfully. Additionally, because each indication group comprised many different indications, the rationale for using it as a predictor variable was weakened in comparison to using the specific indication (which we knew in advance would not be possible given EPV considerations), as indications within indication groups could differ considerably. Indication group was therefore removed.



	RandD	EU.VC.Companies	EU.VC.M.EUR	AROM	XLogP	Distinct.Companies	Molecular.Weight	RBC	HBDC	HBAC	TPSA	Citations	Submitted.Patents	Granted.Patents	US.Interest	UK.Interest	EA.GDP	US.GDP
US.GDP																		1
EA.GDP																	1	0.6
UK.Interest																1	0.46	0.16
US.Interest															1	0.73	0.7	0.34
Granted.Patents														1	-0.02	-0.02	0	-0.02
Submitted.Patents													1	0.91	0.01	0	0.03	0
Citations												1	0.17	0.2	-0.01	-0.02	0.01	0.02
TPSA											1	-0.01	-0.07	-0.08	0	0.02	-0.01	-0.01
HBAC										1	0.84	-0.04	-0.08	-0.09	-0.05	-0.03	-0.05	-0.04
HBDC									1	0.58	0.79	-0.01	-0.06	-0.07	0.03	0.02	0	0.02
RBC								1	0.44	0.57	0.58	-0.08	-0.07	-0.07	0.05	0.04	0.03	0.05
Molecular.Weight							1	0.74	0.51	0.77	0.73	-0.11	-0.11	-0.13	-0.02	-0.01	-0.03	-0.01
Distinct.Companies						1	0.01	-0.01	0.03	0.03	0.04	-0.02	-0.02	-0.01	-0.01	-0.01	0.02	0.01
XLogP					1	-0.01	0.3	0.26	-0.32	-0.13	-0.29	-0.08	-0.03	-0.06	0	-0.02	0.03	0.03
AROM				1	0.37	0.02	0.32	0.11	-0.08	0.16	0.07	-0.09	-0.12	-0.13	-0.08	-0.1	-0.02	-0.01
EU.VC.M.EUR			1	0.08	0.07	0.04	0.07	0.04	-0.01	0.08	0.04	0.02	0.09	0.07	-0.07	-0.28	0.24	-0.04
EU.VC.Companies		1	0.9	0.1	0.07	0.07	0.06	0.02	-0.02	0.07	0.02	0.02	0.09	0.09	-0.28	-0.38	0.09	-0.04
RandD	1	0.68	0.61	0.16	0.05	0.05	0.05	-0.02	-0.05	0.08	0.02	0.03	0.01	0.03	-0.6	-0.65	-0.22	-0.3

Figure 5.4: Correlation matrix for all discovery dataset numerical variables

Pearson's correlation coefficients for all pairs of numerical candidate predictor variables for which data were collected across all drugs. Abbreviations: AROM = number of aromatic rings, EA GDP = European Area gross domestic product change (annual %), EU VC Companies = number of companies receiving venture capital in EU (10s of companies) HBAC = hydrogen bond acceptor count HBDC = hydrogen bond donor count, RandD = pharma R&D spend, RBC = rotatable bond count, TPSA = topological polar surface area (Å^2), US GDP = United

States gross domestic product change (annual %), XLogP = logarithm of the octanol/water partition coefficient (XLogP method). For full descriptions see Table 5.1 and methods.

To assess collinearity, a correlation matrix was created for all numerical candidate predictors (Figure 5.4) and one of each pair of the most highly correlated variables removed (apart from physicochemical parameters, as discussed above). The number of submitted and granted patents were highly correlated ($r = 0.91$) so granted patents were removed as a predictor. Granted patents were removed because it is likely that more data on submitted patents would be available at the time of prediction given the timelines for patents to be granted (generally at least four years from submission) compared to publication of submitted patent applications (generally 18 months from submission)[157].

In the economic variables, UK interest was correlated with US interest ($r = 0.73$), and the amount of venture capital investment was highly correlated with the number of early-stage companies receiving investments ($r = 0.90$), so one of each was removed (UK interest and amount of venture capital investment). EPV prior to commencement of modelling was therefore 10.6.

5.5.1.2 Variable Selection During Modelling

No variable selection during modelling was planned. However, when models were fit with the ROA predictor variables, we found that their association with approval was inflated because the data were not limited to that available at the date at which the prediction was made (see Results, Table 5.2). Rather, they had been updated in Cortellis for each drug throughout its lifecycle, and approved drugs were therefore much more likely to have a listed ROA than discontinued drugs. The ROA predictors were therefore removed.

5.5.1.3 Final Predictor Variables

The final predictor variables for the discovery model are listed in Table 5.1.

5.5.1.4 Dataset Characteristics

The summary statistics for each predictor are highly similar for approved and discontinued drugs, apart from ROA (Table 5.2) which was removed as discussed above (Section 5.5.1.2). The distribution of variables was also similar (examples in Supplementary Figure 8.2). Start year for drugs in the final dataset ranged from 1967 to 2014 (median 1998, IQR [1996, 2003]) and was similar for approved and discontinued drugs (Figure 5.5).

Table 5.2: Descriptive statistics for discovery dataset

For information on missing data for each variable, see table footnotes. Data presented are after estimation of any missing data. In addition to the missing data discussed specifically here, some drugs were excluded on the basis of lack of availability of any physicochemical information due to an inability to match them with CIDs (Figure 5.3). See sensitivity analysis in Section 5.5.1.6.1 and Limitations, Section 5.6.4 for further information and discussion of this.

Predictor	**All (n = 1,220)**	**Approved (n = 212)**	**Discontinued (n = 1,008)**
Developer Characteristics			
Academic origin (%)	67 (5.5)	15 (7.1)	52 (5.2)
Non-originator company developing drug (%)	280 (23.0)	64 (30.2)	216 (21.4)

Predictor	**All** **(n = 1,220)**	**Approved** **(n = 212)**	**Discontinued** **(n = 1,008)**
Mean number of companies developing drug (SD)	1.1 (0.3)	1.1 (0.3)	1.1 (0.3)
Route of Administration†			
Oral ROA (%)	536 (43.9)	168 (79.2)	368 (36.5)
No ROA provided (%)	534 (43.8)	11 (5.2)	523 (51.9)
Other ROA (%)	172 (14.1)	42 (19.8)	130 (12.9)
Physicochemical Parameters			
Mean molecular weight, g/mol (SD)	442 (175)	434 (168)	443 (177)
Mean hydrogen bond donor count (SD)	2.2 (2.0)	2.0 (1.7)	2.2 (2.1)
Mean hydrogen bond acceptor count	6.4 (3.6)	6.8 (3.6)	6.3 (3.5)
Mean rotatable bond count	7.0 (5.2)	6.6 (4.6)	7.0 (5.3)
Mean topological polar surface area, Å^2 (SD)	104 (64)	104 (60)	104 (65)
Mean number of aromatic rings (SD)	2.2 (1.3)	2.2 (1.4)	2.2 (1.3)
Mean partition coefficient, XLogP (SD)	3.5 (2.6)	3.3 (2.4)	3.5 (2.6)
Literature			
Mean citations (SD)	1.1 (10.2)	2.4 (17.9)	0.8 (7.6)
Mean submitted patents (SD)	24 (286)	40 (294)	20 (285)
Economic			
Mean pharma R&D spend, 100M USD, adjusted to 2015 (SD)‡	320 (110)	307 (116)	323 (109)
Mean US interest rate (SD)	4.4 (1.8)	4.3 (2.0)	4.4 (1.7)
Mean number of companies receiving venture capital in EU, 10s of companies§ (SD)	171 (100)	168 (107)	171 (99)
Mean EA GDP change (SD)	1.8 (1.3)	1.6 (1.4)	1.8 (1.2)
Mean US GDP change (SD)	2.1 (1.3)	2.0 (1.3)	2.1 (1.3)

† Removed during modelling due to confounding.

‡ Data was missing for one drug with a start date of 1967. 1970 value was instead used.

§Data prior to 1996 was not available (245 drugs [181 discontinued, 64 approved]). These were estimated as described in Section 5.3.5.5.1. Prior to estimation of historical data, mean (SD) of approved and discontinued drugs was 209 (104) and 193 (97), respectively.

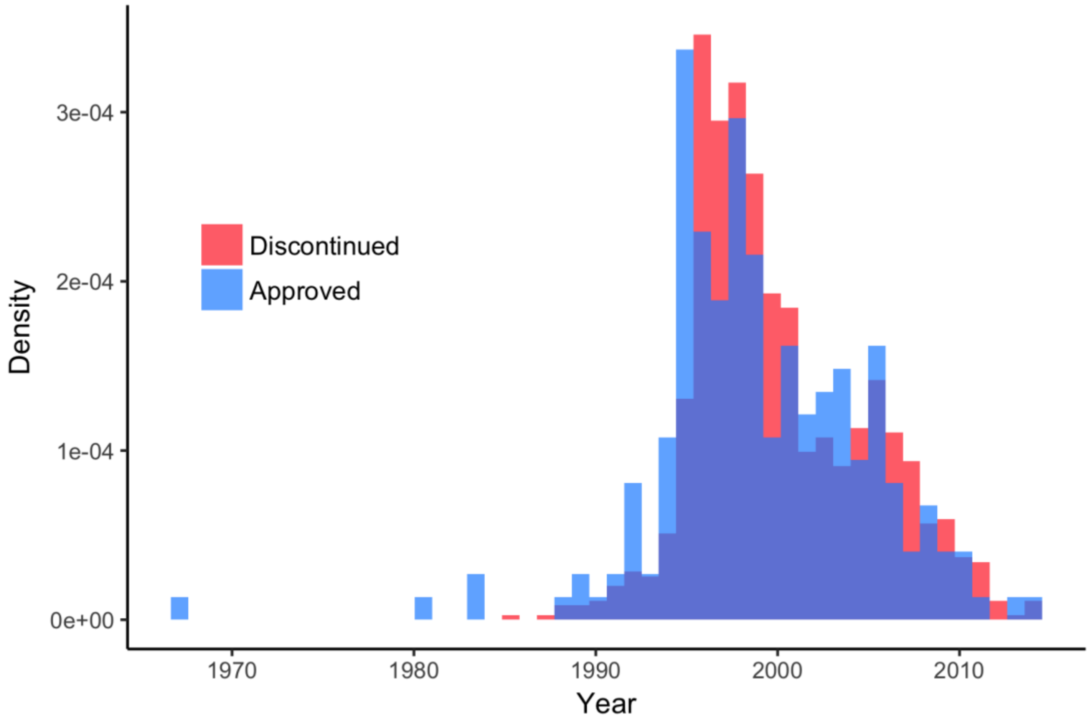


Figure 5.5: Start year for drugs included in the discovery dataset

Start year was comparable for approved (n = 212) and discontinued (n = 1,008) drugs. Median start year (IQR) was 1998 (1995, 2003) and 1998 (1996, 2003), respectively.

5.5.1.5 Logistic Regression Model

A logistic regression model was fitted to the entire dataset using a full model approach. The assumption of linearity in the logit was checked graphically for the linear model and was in most cases satisfied (Supplementary Figure 8.3). Modelling the predictors as linear gave an acceptable fit based on visual inspection of the deviance residuals (Supplementary Figure 8.4), and we therefore used a linear model as the final model.

An apparent AUC performance estimate of 0.63 was obtained, and after correction for model optimism by bootstrapping was estimated at 0.59 (95% CI [0.55, 0.63]). The full model specification is shown in Table 5.3. Apparent and bias-corrected calibration in the discovery dataset was assessed graphically *via* bootstrapping (Figure 5.6). Outside of predicted probabilities 0.1 to 0.3, the model does not appear to be well calibrated, probably because there are very few predictions (Supplementary Figure 8.5).

Table 5.3: Logistic regression coefficients for discovery model with predictors sorted by decreasing absolute Z-value

Intercept and Predictors*	Coefficient	SE	Z-Value
Intercept	-0.0768	0.753	-0.102
Hydrogen bond acceptor count	0.1598	0.046	3.448
Pharma R&D spend, 100M USD, adjusted to 2015	-0.0040	0.001	-2.891
Number of companies receiving venture capital in EU, 10s of companies	0.0023	0.001	2.057
Company change (change = 1)	0.4199	0.208	2.018
Molecular weight, g/mol	-0.0019	0.001	-1.611
Citations	0.0099	0.007	1.419
Hydrogen bond donor count	-0.0701	0.065	-1.077
EA gross domestic product change	-0.1059	0.109	-0.974
US gross domestic product change	-0.0662	0.080	-0.824
Number of aromatic rings	0.0387	0.072	0.537
US interest rate	-0.0425	0.083	-0.513
Topological polar surface area, Å^2	-0.0017	0.004	-0.482
Submitted patents	0.0001	0.000	0.468
Academic commercial (commercial = 1)	-0.1488	0.318	-0.467

Intercept and Predictors*	Coefficient	SE	Z-Value
Rotatable bond count	-0.0111	0.025	-0.447
Partition coefficient, XLogP	0.0137	0.055	0.251
Distinct companies	-0.0263	0.278	-0.094

** See Table 5.1 and methods for further information on predictor variables. Approval was represented by the outcome '1' during modelling. Therefore, higher predicted probabilities correspond to a higher probability of approval.*

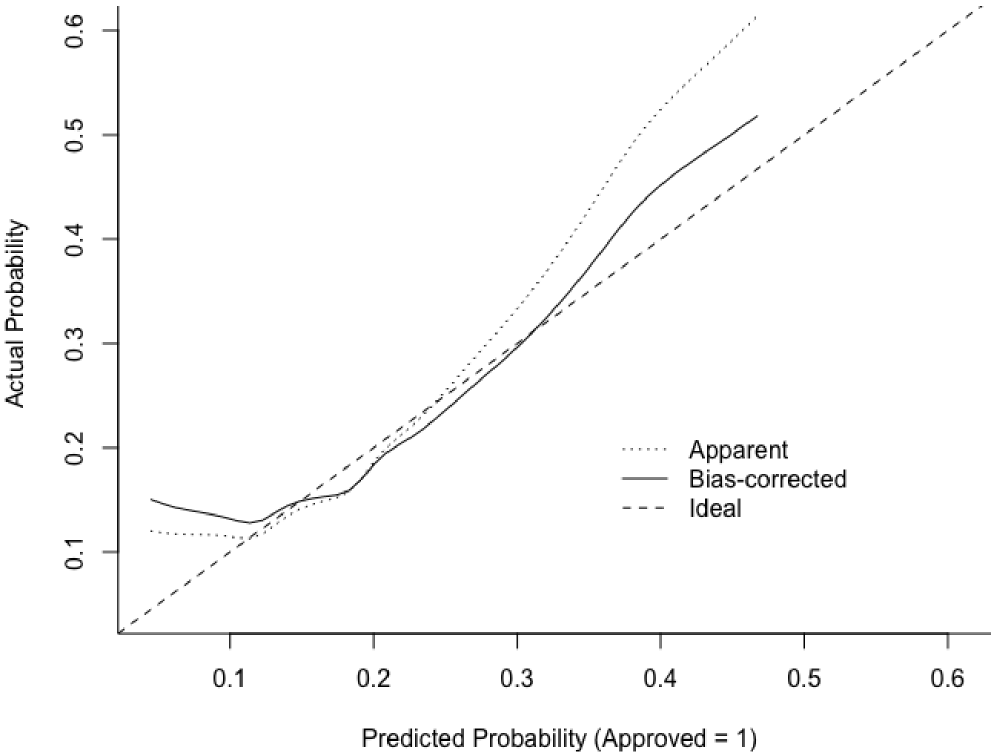


Figure 5.6: Apparent and bias-corrected calibration plot for discovery model

The discovery model was used to generate predicted probabilities for the discovery dataset which are plotted against actual probabilities using a lowess smoother to give the apparent calibration. Bootstrapping (1,000 resamples) was used to estimate the optimism in the apparent calibration and bias-corrected calibration is plotted. Perfect agreement between *observed and predicted probabilities is represented by the ideal line. Between 0.1 and 0.3 the model is well calibrated; however, outside of this range calibration is poor, probably because there are very few predictions (Supplementary Figure 8.5)*

5.5.1.6 Sensitivity Analyses

5.5.1.6.1 Dealing with Sampling Bias

A sensitivity analysis of the impact of changing the class distribution in the test dataset was evaluated because the true distribution of approved and discontinued drugs is not known. The performance across test datasets with different proportions of approved drugs was very similar to the apparent performance of the physicochemical model (Table 5.4).

Table 5.4: Performance of discovery model on full discovery dataset and sensitivity analysis with undersampled approved drugs

ID*	Test Set	Approved (as % of Dataset)	Discontinued (as % of Dataset)	AUC (95% CI)
1	All	17.4%	82.6%	0.63 (NA)
2	Undersample approved	15.0%	85.0%	0.63 (0.61-0.65)
3	Undersample approved	12.5%	87.5%	0.63 (0.60-0.66)
4	Undersample approved	10.0%	90.0%	0.63 (0.59-0.67)
5	Undersample approved	7.5%	92.5%	0.63 (0.58-0.68)

** ID 1 is the apparent performance. IDs 2-5 follow the methodology detailed in Section 5.3.7.4.1.*

5.5.1.6.2 Outliers and Influential Observations

A Bonferroni outlier test did not identify any significant outliers. Graphical examination of Cook's distance identified two observations that were particularly influential (272 and 569; Figure 5.7), and the impact of removing these observations on the regression coefficients was therefore investigated (Supplementary Table 8.8). Manual review did not reveal any justification for excluding the influential observations, though their inclusion did alter some coefficient estimates considerably.

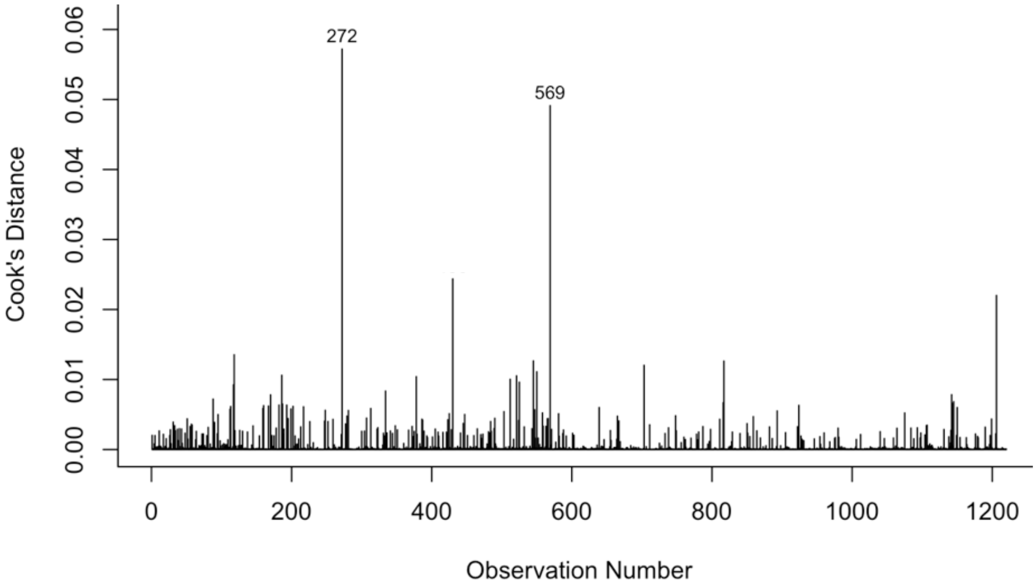


Figure 5.7: Cook's distance for observations in the discovery dataset shows two particularly influential observations

Cook's distance is a measure of the change in coefficient estimates with removal of an observation and was calculated for each observation in the discovery dataset. Two particularly influential observation (272 and 559) were identified for further analysis (Supplementary Table 8.8).

5.5.1.6.3 Impact of Non-physicochemical Predictors

An additional analysis with only the physicochemical predictors was conducted to investigate the improvement in performance provided by the other predictors which resulted in an optimism corrected AUC of 0.55 (95% CI [0.51, 0.59]).

5.5.2 Physicochemical Model

5.5.2.1 Dataset Characteristics

As with the discovery dataset, the summary statistics and distributions for each predictor are highly similar for both approved and discontinued drugs in the physicochemical dataset (Table 5.5, Supplementary Figure 8.7).

Table 5.5: Descriptive statistics for physicochemical dataset

No variables had missing data, though some drugs were excluded on the basis of lack of availability of any physicochemical information due to an inability to match them with CIDs (Figure 5.3). See sensitivity analysis in Section 5.5.2.5.1 and Limitations, Section 5.6.4 for more information and discussion of this.

Predictor	All Drugs (n = 3,229)	Approved Drugs (n = 904)	Discontinued Drugs (n = 2,325)
Mean molecular weight, g/mol (SD)	429 (185)	417 (186)	434 (185)
Mean hydrogen bond donor count (SD)	2.2 (2.3)	2.3 (2.3)	2.1 (2.3)
Mean hydrogen bond acceptor count (SD)	6.3 (3.8)	6.6 (4.0)	6.1 (3.7)
Mean rotatable bond count (SD)	6.8 (5.4)	6.4 (4.7)	6.9 (5.7)
Mean topological polar surface area, Å^2 (SD)	102 (73)	107 (70)	101 (74)
Mean number of aromatic rings	2.1 (1.3)	1.8 (1.3)	2.2 (1.3)

Predictor	All Drugs (n = 3,229)	Approved Drugs (n = 904)	Discontinued Drugs (n = 2,325)
(SD)			
Mean partition coefficient, XLogP (SD)	3.2 (2.8)	2.6 (2.9)	3.5 (2.7)

5.5.2.2 Variable Selection

No variable selection prior to modelling was required because the sample size was sufficiently large and the number of predictors small, and checks for collinearity showed acceptable correlation (*r* < 0.9 in all cases; Supplementary Figure 8.6). No variable selection during modelling was performed. The final predictor variables were therefore the same as the candidate predictor variables, and the same as the physicochemical predictors included in the discovery dataset.

5.5.2.3 Logistic Regression Physicochemical Model

As with the discovery dataset, a full model approach was taken to fit a logistic regression model to the physicochemical dataset with all predictors modelled as linear. The assumption of linearity in the logit was checked visually (Supplementary Figure 8.8) and was satisfied, and the fit based on the deviance residuals was acceptable (Supplementary Figure 8.9).

The apparent performance in terms of AUC was 0.63 and when corrected for optimism by bootstrapping was 0.62 (95% CI [0.60, 0.64]), an improvement on the discovery dataset. The full physicochemical model specification is presented in Table 5.6. Within this dataset, the model is reasonably well calibrated over a larger range than the discovery model (Figure 5.8), though there are again very few extreme predictions

(Supplementary Figure 8.10), resulting in poorer calibration at very low or high probabilities.

Table 5.6: Logistic regression coefficients for physicochemical model with predictors sorted by decreasing absolute Z-value

Intercept and Predictors*	Coefficient	SE	Z-Value
Intercept	-0.4030	0.111	-3.621
Number of aromatic rings	-0.2162	0.037	-5.832
Hydrogen bond acceptor count	0.1125	0.025	4.443
Octanol/ water partition coefficient, XLogP	-0.0555	0.027	-2.032
Rotatable bond count	-0.0194	0.012	-1.638
Topological polar surface area, Å^2	-0.0021	0.002	-1.174
Molecular weight, g/mol	-0.0007	0.001	-1.120
Hydrogen bond donor count	-0.0116	0.031	-0.370

** See Table 5.1 and methods for further information on predictor variables. Approval was represented by the outcome ‘1’ during modelling. Therefore, higher predicted probabilities correspond to a higher probability of approval.*

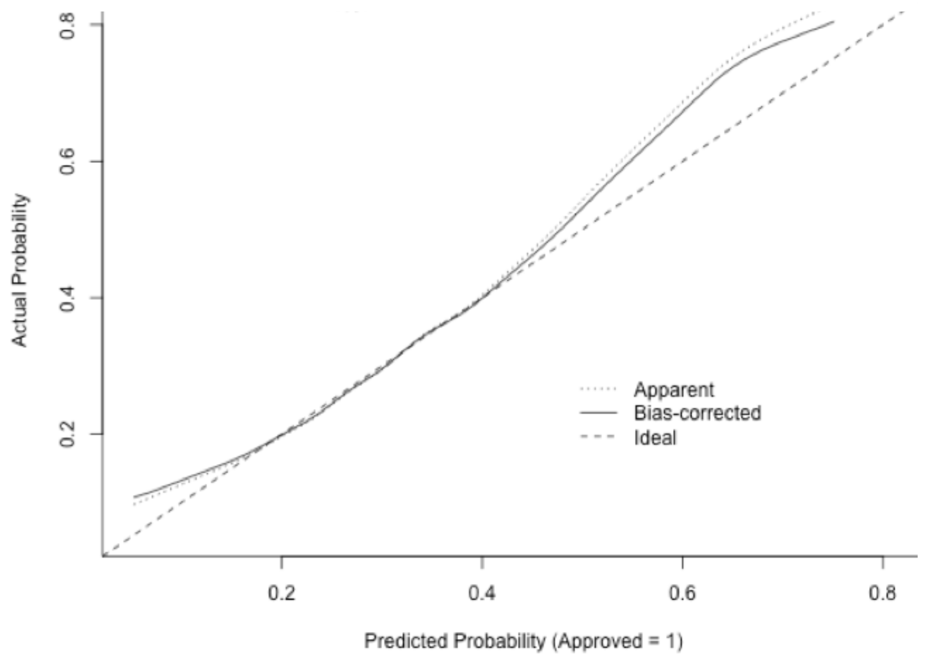


Figure 5.8: Apparent and bias-corrected calibration plot for physicochemical model

The physicochemical model was used to generate predicted probabilities on the physicochemical dataset which are plotted against actual probabilities using a lowess smoother to give the apparent calibration. Bootstrapping (1,000 resamples) was used to estimate the optimism in the apparent calibration and bias-corrected calibration is plotted. Perfect agreement between observed and predicted probabilities is represented by the ideal line. The model is well calibrated over a larger range (between 0.2 and 0.5) than the discovery model; however, outside of this range calibration is again poor, probably because there are very few predictions (Supplementary Figure 8.10).

5.5.2.4 Rule-of-Five Comparison

The physicochemical model performs considerably better than Lipinski's Ro5 in this dataset (Figure 5.9). The Ro5 provides almost no ability to discriminate between the approved and discontinued drugs in our dataset.

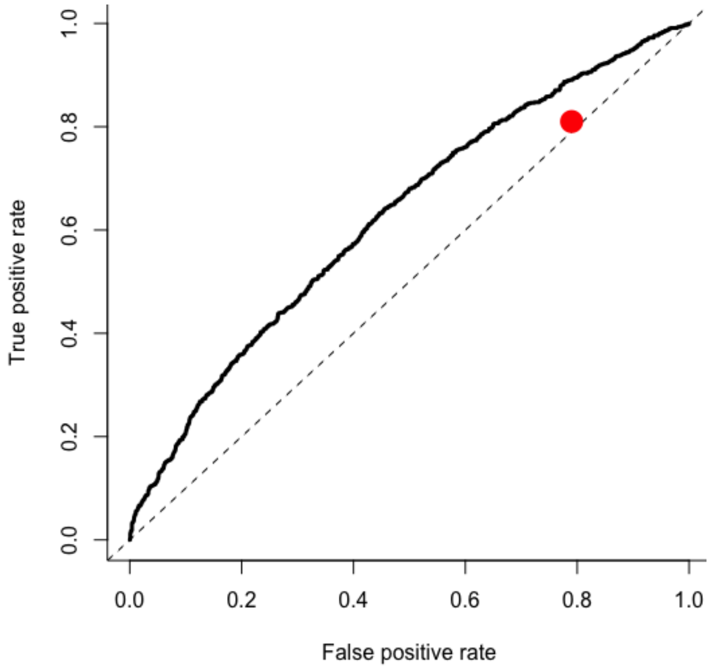


Figure 5.9: ROC curve comparing physicochemical model to Lipinski's Ro5

The black line is the apparent receiver operating characteristics (ROC) curve of the physicochemical model, the dashed line represents a classifier with no ability to discriminate, and the red dot is the performance of Lipinksi's Ro5 (extended) in our dataset, which provides almost no ability to distinguish between approved and discontinued drugs. To be considered Lipinski compliant, molecules could have no more than one violation of the following: molecular weight ≤ 500 Daltons, oil/water distribution coefficient (LogP) is ≤ 5, H-bond donor count ≤ 5, H-bond acceptor count ≤ 10, rotatable bond count ≤ 10.

5.5.2.5 Sensitivity Analyses

5.5.2.5.1 Dealing with Sampling Bias

Sensitivity analyses showed that the model performance was relatively robust in terms of performance to changes in the proportion of approved drugs in the test set, and in both the training and test set (Table 5.7).

Table 5.7: Performance of physicochemical model on full physicochemical dataset and sensitivity analysis with undersampled approved drugs

ID*	Training Set	Test Set	Approved (as % of Dataset)	Discontinued (as % of Dataset)	Corrected AUC (95% CI)
1	All	All	28.0%	72.0%	0.63 (NA)
2	All	Undersample approved	20.0%	80.0%	0.63 (0.61-0.64)
3	All	Undersample approved	15.0%	85.0%	0.63 (0.61-0.65)
4	All	Undersample approved	10.0%	90.0%	0.63 (0.60-0.66)
5	All	Undersample approved	5.0%	95.0%	0.63 (0.58-0.67)
6	All (bootstrapped)	All	28.0%	72.0%	0.62 (0.60-0.64)
7	Undersample approved (bootstrapped)	Undersample approved	20.0%	80.0%	0.62 (0.60-0.65)
8	Undersample approved (bootstrapped)	Undersample approved	15.0%	85.0%	0.62 (0.59-0.65)
9	Undersample approved (bootstrapped)	Undersample approved	10.0%	90.0%	0.62 (0.59-0.66)
10	Undersample approved (bootstrapped)	Undersample approved	5.0%	95.0%	0.62 (0.57-0.67)

** ID 1 is the apparent performance. IDs 2-5 follow the methodology detailed in Section 5.3.7.4.1. ID 6 follows the internal validation procedure for the full dataset described in Section 5.3.7.3. IDs 6-9 follow the methodology detailed in Section 5.5.2.5.1.*

5.5.2.5.2 Outliers and Influential Observations

A Bonferroni outlier test did not identify any significant outliers. Graphical examination of Cook's distance identified one observation with considerably higher influence than all others (Figure 5.10; observation 666), and the impact of removing this observation on the regression coefficients was therefore investigated

(Supplementary Table 8.9). Manual review of the observation did not reveal any justification for excluding it, though some coefficients did vary considerably.

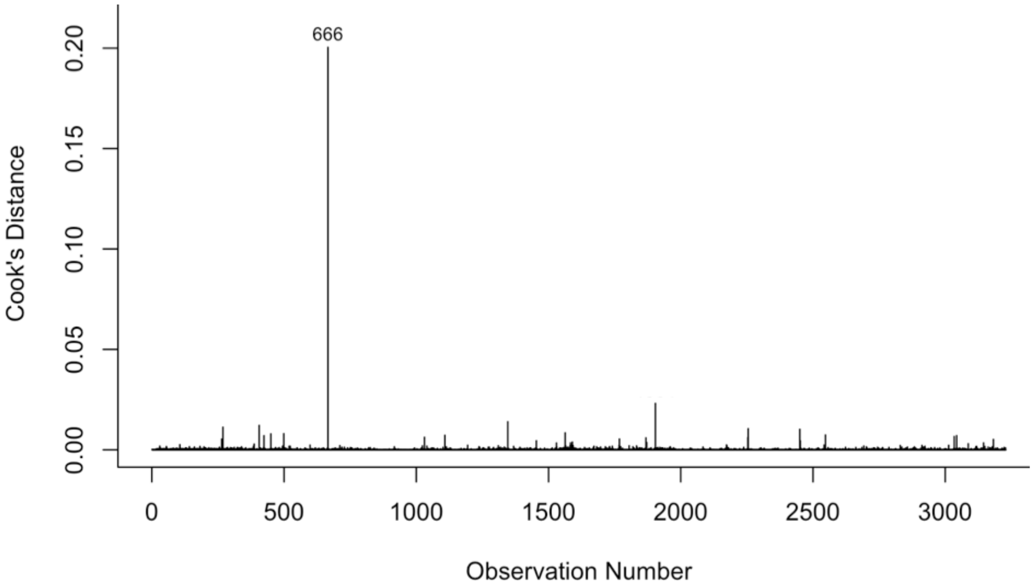


Figure 5.10: Cook's distance shows an observation with large influence in the physicochemical dataset

Cook's distance is a measure of the change in coefficient estimates with removal of an observation and was calculated for each observation in the discovery dataset. One particularly influential observation (666) was identified for further analysis (Supplementary Table 8.9).

5.6 Discussion

Two logistic regression models of small molecule regulatory approval were developed from two datasets of approved and discontinued drugs: one smaller dataset incorporating a broad range of predictors, and one larger dataset containing only physicochemical parameters. Estimates of performance obtained in both datasets were similar, with the physicochemical model performing slightly better (optimism adjusted

AUC = 0.62, 95% CI [0.60, 0.64]). To our knowledge, this work represents the first attempt to develop a valid predictive model of regulatory approval based on analysis of approved and discontinued drugs that could be applied at the early-stages of development. This work also adds to the drug-likeness literature and the physicochemical model performed considerably better than Lipinski's Ro5 in this dataset.

5.6.1 Comparison to Other Models of Regulatory Approval

Unlike other models to predict regulatory approval, we attempted to make predictions from an earlier stage in development and developed a model only applicable to small molecules. A quantitative comparison of the performance of our model to other models was therefore not possible. DiMasi *et al.*[125] developed a tool for predicting regulatory approval of anti-cancer drugs, though the sample size for developing the model was very small (62 drugs) and there was no method of internal validation. Lo *et al.*[127] developed a model that performed significantly better when compared to DiMasi *et al*. They attempted to predict approval across a wide range of therapeutic approaches and indications using a wide range of predictor variables, from both phase II (AUC = 0.78, "P2APP model") and phase III (AUC = 0.81, "P3APP model") trial completion. They attempt to predict approval of drug-indication pairs, rather than unique drugs for any indication, and the phase II model is developed on the largest database used in any predictive model of regulatory approval (4,073 drugs)[d].

Despite the large dataset, apparently robust methodology and high predictive performance, it is interesting to note that by far the two most informative predictors in both the P2APP and P3APP models are i) that the trial is completed with the positive outcome or primary endpoint(s) met, and ii) whether the trial was completed or terminated. It is logical that if a trial is not completed and/or its primary outcomes are not met, its probability of progressing to the next stage of development will be much lower than the converse, and it is likely that anyone familiar with drug development would be able to identify drugs unlikely to progress based on these predictors alone. Developing the models only on the subsets of data that did meet these criteria, or only those that were selected to advance to the next stage, might have provided a more useful assessment of the marginal benefit of the models. Alternatively, the univariate associations between these predictors and the outcome, or the summary statistics between the different groups, would allow the reader to assess more directly the explanatory power of these variables, though neither of these is available. DiMasi *et al.* use a similar predictor to meeting primary outcomes (termed “activity”) in their model, which is by far the most predictive variable. Our discovery dataset, by contrast, does

---

[d] Our systematic review (Chapter 4) identified a third paper[126] which we do not discuss here because issues in the methods render interpretation of the performance challenging. See Chapter 4 for more information.

not contain any predictor variables that could effectively inflate model performance estimates.

5.6.2 Comparison to Other Physicochemical Models

Many models for identifying drug-like compounds based on physicochemical descriptors exist[143,145,150,181–184], though to our knowledge our model is the first attempt to explicitly predict approval or discontinuation of drugs on the basis of physicochemical descriptors, rather than to describe drug-likeness more generally.

Many models that do exist result in the presentation of simple “rules” for determining whether or not a compound is drug-like. This involves categorisation of continuous variables according to cut-off points defined from the development dataset, which is known to cause a range of issues and is rarely appropriate[95,138]. As an example, consider Lipinski’s Ro5 described in Section 5.4.4.1: the use of a cut-off for molecular weight of 500 Da implies two drugs with molecular weights of 499 and 501 Da are as dissimilar as two drugs of 200 and 1,000 Da, which is clearly not the case. In the context of drug development, there is no real need for such simplistic rules, since it will always be possible to compute any calculations and any slight increased time requirements would surely be offset by modelling improvements. Additionally, rule-based approaches often generate simply a pass/fail outcome, which means that their utility in comparing compounds is more limited because they can be less easily ranked. It has been shown that 4 of the top 10 selling US drugs in 2010 would not have passed one commonly used rule[185] (the 3/75 rule[184]) and it is therefore clear that in some circumstances the utility of these rules is limited. Although Lipinski’s rules are not necessarily intended for distinguishing approved vs. failed drugs (the rules were originally derived from a set of Phase II drugs, some of which probably eventually failed), we provide a comparison and show that our physicochemical model exhibits considerably better performance (Figure 5.9).

Recognising the issues with rule-based approaches, Bickerton *et al.* developed a quantitative measure of drug-likeness based on analysis of approved drugs, from which they derive a “desirability function” based on the physicochemical parameters listed in Table 5.1. However, the evaluation of this model is flawed. The performance is estimated in a set of compounds including 72% of those drugs on which the model is developed and a large number of small molecule ligands. No procedure to account for the use of the same data in the training and test set is used (e.g. bootstrapping), and 475 compounds that were structurally similar to the approved drugs included in the test set were removed to “prevent ambiguity”, which is likely to further inflate performance estimates. It should be noted, however, that this model is not directly comparable to ours because it focusses specifically on orally administered small molecule drugs. Other approaches to develop probabilistic models to distinguish drugs and non-drugs have been attempted by García-Sosa *et al.* and Yosipof *et al.*; an issue in these papers, which are each based on the same data, is that balanced samples of approved drugs and non-drugs are used to train and test the model, and no assessment of the performance with more representative outcome ratios is provided.

Regardless, in the majority of this literature, the comparators (i.e. the negative control) are compounds from large libraries of small molecules. Our approach differs in that we use discontinued drugs as the comparator, which is likely to represent a more difficult classification problem because the parameter distributions are so similar[186] (Supplementary Figure 8.7). We combine the approaches taken to predict regulatory approval (i.e. the use of approved and discontinued drugs for modelling) with the use of predictors commonly used in assessment of drug-likeness, which we do not think has been done previously. The physicochemical model also represents the largest analysis of physicochemical parameters of approved and discontinued small molecules that we have identified.

5.6.3 Interpreting Performance

An AUC of 0.62 would often be considered poor and compared to, for example, the P2APP AUC of 0.78 reported by Lo *et al.* it appears low. However, the AUC must be taken in the context of the classification problem. A model that can be used prior to clinical trials can potentially identify drugs likely to fail and save the full clinical costs of those drugs; however, its use will inevitably result in some drugs not being progressed that would have eventually been approved (false negatives), thus resulting in a large loss of future revenue or patient benefit. When making predictions at phase II, for example, the potential cost savings of not advancing drugs that will fail are lower (since phase I and II have been completed and the cost incurred) but there will be fewer false negatives because some drugs will have already been correctly discontinued due to the results of the phase I trials. It may be that in both cases, with the current levels of performance, neither model is better than no modelling, because the cost of false negatives is so high. To determine the true utility of these predictive models, work is needed to assess their potential impact, both financially and to patients; modelling approaches to do this have been developed and could potentially be adapted for use in this context[187].

Given the low AUC values, the possibility that the model has no real predictive value must also be considered. All predictor variables in both datasets (Table 5.2 and Table 5.5) are very similar between approved and discontinued drugs. Physicochemical properties are already used in filters during drug discovery[117], so it is possible that in the later stages of development there are no longer any systematic differences in these parameters, and that more complex models are therefore required. It is also possible that the physicochemical predictors simply do not have predictive value. Other proposed predictors (developer characteristics, literature, commercial) may also simply not be predictive. This finding would still be useful and would narrow the range of predictors for future work in this area.

5.6.4 Limitations

The samples used for development of both models do not include all potentially relevant compounds. There were 5,233 compounds potentially suitable for inclusion in both of our models, and the final datasets were considerably smaller. The largest number of drugs were removed because drug names could not be matched to CIDs, meaning no physicochemical data were available. Imputation of missing data is generally recommended[95] but was not possible because no data from which the physicochemical data could reasonably be imputed were available, as Cortellis does not contain data on any parameters describing the molecule itself. Therefore, the sample is not fully representative of all potentially relevant data, and in particular, approved drugs were overrepresented. To address this, several sensitivity analyses were performed which demonstrated that the AUC estimates were relatively robust to changes in outcome distribution. Although we examined apparent and bias-corrected calibration in our dataset, in data with different outcome distributions the model calibration would likely be worse and it should therefore not be used to directly estimate the probability of approval (rather than to compare and rank compounds). To address this, if more information about the outcomes was available, the model could be recalibrated[188] and would potentially then be useful for direct estimations of the probability of approval.

Both models included LogP as a predictor. LogP, unlike the other physicochemical parameters included in the models, is an experimentally measured property of a compound that cannot be calculated exactly. It is instead predicted, and there are a number of different methods available for this prediction. We used one commonly used and freely available method, XLogP[153], but very recent comparisons suggest some other methods may be superior[189]. Different methods for calculating LogP could influence the estimated coefficients in our model and the predictive performance, so exploring them in the future could be valuable.

For both models, there is a need for external validation before either should be used for decision-making. Though every effort has been made to quantify the optimism associated with our performance estimates, independent data is needed to confirm this. The DrugBank database contains physicochemical information 2,627 approved small molecule drugs (www.drugbank.ca, accessed 11th September 2018), so there are at least

1,723 approved drugs that were not in our dataset which could be used. Identification of discontinued drugs is less straightforward; other commercial databases may be the best option. Alternatively, the model could be applied to the current drug pipeline and its performance evaluated over time. It would take many years to complete this process, though indications of performance could be gathered now, as Lo *et al.* show[127], and revisited periodically.

5.6.4.1 Discovery Model

The discovery dataset is limited to a large extent by the availability of data. Variables such as ROA were back-filled in the database, resulting in an inability to use them as predictors, and there were too few events to meaningfully include indication as a predictor. While we accurately collected literature information from PubChem, the vast majority of literature data in PubChem is gathered through data mining techniques[151], which may not be completely accurate. Additionally, we estimated the number of companies receiving venture capital for drugs with a start date prior to 1996 (Table 5.2) from number of investments made in that time period. Multiple imputation using chained equations (MICE) is generally the recommended approach for imputing missing data[95]; however, because variance in number of companies in available data could be almost completely explained by number of investments ($R^2 = 0.97$), any improvement with MICE would likely be marginal and we believe not worthwhile given the additional complexity it introduces into analysis.

Despite issues with data availability, there was an improvement in performance in the discovery dataset model built using all parameters (AUC = 0.59, 95% CI [0.55, 0.63]) compared to just physicochemical parameters (AUC = 0.55, 95% CI [0.51, 0.59]).

Therefore, if a larger dataset with all information, or an augmented physicochemical dataset with additional parameters included in the discovery dataset, could be developed, performance improvements could be realised.

5.6.4.2 Physicochemical Model

The physicochemical model encompasses a wide variety of compounds across different ROA and indication. Many other papers, on the other hand, have included solely oral drugs[143,145], or develop different models for different indications[181]. It is possible that our approach of grouping all approved and failed drugs is too simplistic and cannot capture inherent differences between drugs for different indications or ROA. To some extent, this must be true: for example, the risk-benefit in drugs for certain indications will be different across drugs, meaning that the same physicochemical parameters leading to the same side effects in one indication could result in discontinuation, whereas in another they could be acceptable.

Additionally, in our data, the reason for discontinuation of drugs is not known. Approximately 30% of drug discontinuations are thought to be due to “economic” reasons[190], which are independent of the specific physicochemical parameters. In the discovery dataset, we included predictors that could represent the economic aspects of a drug, whereas in the physicochemical model we did not. Ideally, therefore, we would have included in the physicochemical dataset only those drugs known to have failed for safety and/or efficacy. This information is not available in our data, so this could not be done.

Our approach did, however, allow a large sample size to be used, which might offset the benefits of creating different models for different drug groups, or of developing a dataset with drugs with only known reasons for discontinuation, which would be considerably smaller. Although these considerations may place an upper limit on the performance of our approach, internal validation has shown that it is, at least to some extent, predictive and therefore potentially useful. Future work with enhanced datasets could explore the trade-off between these issues further.

5.6.5 Outstanding Work

There are several analyses planned as part of this project which could not be completed within the timescale of the DPhil. In developing the physicochemical dataset, there were 339 drug names that matched to multiple CIDs, from which a random CID was selected for each name. Although we think it is unlikely that selecting different CIDs will influence the results, sensitivity analysis is planned to check this by using different random samples to generate models and compare the coefficients and model performance.

We compared our model to Lipinksi's Ro5 because this is one of the most well-known and commonly used physicochemical filters, though there are other comparators that are also relevant. In particular, we hoped to compare our model to the quantitative estimate of drug-likeness[145], but were unable to because the method for calculation of number of structural alerts could not be clearly determined. Work is currently planned to generate at least a similar variable which can be used in a model update and subsequent comparison. There are at least two other relevant comparators[150,182] for the physicochemical model that could be calculated simply from our data[e]. It should be stressed that, like the Ro5, these rules are not aimed specifically at distinguishing approved from discontinued drugs and are generally used earlier in drug discovery.

5.6.6 Using the Model

Notwithstanding the limitations and outstanding work, readers may wish to apply or attempt to externally validate the work presented here. The optimism corrected performance of the physicochemical model is better than that of the discovery model so the physicochemical model (Table 5.6) should be used. This model is applicable to drugs at any stage of development covered by the Cortellis database, from preclinical to clinical. However, given the increasing availability of data in later stages of development that may provide improved evidence relating to the likelihood of approval, the model may be better suited for use early in development when a relatively large number of compounds are being evaluated. It should be stressed that the model may not be suitable for distinguishing compounds likely to be approved from general compound libraries, because those libraries may include compounds dissimilar to those used in our set of discontinued drugs. This model is only suitable for small molecules within the physicochemical parameter space described in Table 5.5.

---

[e] The 4/400 rule[182] suggests that for good absorption, distribution, metabolism and excretion properties molecules should have a molecular weight < 400 and ALogP < 4. Veber *et al.*[150] suggested that compounds with < 10 RBC and TPSA ≤ 140 Å2 (or 12 or fewer hydrogen donors and acceptors) had increased oral bioavailability.

We have evaluated the impact of changes in class imbalance on the AUC of our models and showed that it is relatively robust to these changes, which means that the model retains its ability discriminate between classes. However, it is unlikely that the model in its current form is properly calibrated because the outcome distributions in the development data influence the model's calculated probabilities and bias was introduced during data collection. Therefore, rather than using the output of the model to assess the probability of approval directly, it will be more useful in ranking and prioritising molecules, unless it is recalibrated as mentioned above.

It is common to assign a threshold to the predicted probabilities from a model, above which the outcome is considered positive and below which the outcome is considered negative. The value of this threshold can be determined based on the acceptable balance between the sensitivity and specificity of the model, which differs depending on the context in which the model is used. For our models, the threshold could be determined based on assessments of the impact of the use of the model under different circumstances (discussed above in Section 5.6.3). For similar reasons to the criticisms raised about rules for drug-likeness (loss of information[95]), this is generally not recommended, but in some contexts, greater simplicity for the model user may be required to encourage model use and the use of thresholds could be justified. In the absence of impact assessments and recalibration, using the model for ranking molecules is likely preferable.

5.6.7 Conclusion

We use simple predictor variables to achieve a modest predictive performance for regulatory approval vs. discontinuation of small molecule drugs. Although work is required to validate the model externally and assess its utility in a real-world setting, there is little cost in running this model on a portfolio of small molecule candidates and subsequently tracking their performance over time, thus providing a truly independent model validation. As well as improved quality of data, other improvements to the models are plausible and some are discussed in the subsequent chapter.

6 Discussion

6.1 Introduction

The first part of this chapter discusses some general considerations, limitations and potential future work for topics related to different aspects of the thesis that could not be included in individual chapters. Subsequently, several key themes are summarised.

6.2 Reporting Quality

The main focus of reporting quality in this thesis was patent landscaping (Chapters 2 and 3); however, poor reporting quality was also a significant finding in the context of predicting approval (Chapter 4). Here, we review the reporting quality and reporting guideline literature and make several recommendations. With the benefit of hindsight, several ideas have emerged that were not considered while the research was being conducted but which may be useful for others conducting research in this area.

There is now an exceptionally large body of evidence showing that reporting quality is suboptimal across health science, a large amount of which is generated from systematic reviews[31]. In parallel, there has been a proliferation of reporting guidelines aimed at addressing the identified deficiencies[86] (as of 3rd September 2018, 404 reporting guidelines were listed on the EQUATOR network website [equator-network.org]). The conduct of systematic reviews and development of reporting guidelines use resources that could, presumably, be used to study other biomedical questions. Their utility and efficiency should therefore be examined.

The vast majority of systematic reviews conducted on reporting quality appear to show that reporting quality is insufficient. For example, a major Cochrane review of more than 16,600 randomised controlled trials (RCTs) showed that reporting is suboptimal[59]. In five out of six full texts included in a review of evaluations of reporting quality of network meta-analyses, it was concluded that reporting was not adequate and needed improvement[191]. Four assessments of the quality of RCTs in different dental specialties, published within two years of each other, concluded that reporting quality was poor[192–195]. Indeed, even systematic reviews of reporting guidelines, which are often based on systematic reviews of reporting quality themselves, show that reporting quality is poor in the reporting guidelines[86,196].

Given this evidence, the marginal benefit of additional systematic reviews of reporting quality is likely to be low in the absence of some intervention that might influence reporting quality and warrant evaluation, particularly in areas where the issue is already well characterised, such as RCTs. Evidence in Chapter 4 indicated that, at least in this context, machine learning models were poorly reported and this could represent an interesting avenue for investigation. Regardless, specific rationale for systematic reviews of reporting quality should be clearly elucidated.

6.2.1 Sample Size Considerations

If there is a clear rationale for investigating reporting quality, conducting a systematic review in which all relevant articles are analysed may not be necessary. Systematic reviews became prominent in health research to allow the synthesis of randomised controlled trials[197]. In this context, the inclusion of all relevant studies in the analysis is important; different sample sizes in different studies, for example, will result in different weighting of those studies in statistical analysis, thus resulting in non-uniform contributions to the final estimates of effect size. In systematic reviews aiming to identify research in a particular field (for example, for competitive intelligence), identification of all relevant studies is important because missing one study (for example, the one study already reporting your particular idea) could dramatically change the conclusions. In the context of reporting quality, however, neither of these scenarios usually occurs. Most systematic reviews of reporting quality assign each article included in analysis the same weight, and only the aggregated results are important rather than any individual paper. In this case, parameter estimates from a random sample, assuming it is large enough, will approximate the true parameter values (i.e. those of all relevant articles). Identifying and analysing a random sample of relevant results might therefore be sufficient and more resource efficient. This approach is taken in some reporting quality analyses and should be encouraged[198].

6.2.2 Impact of Reporting Guidelines

Evidence suggests that introduction and endorsement of reporting guidelines generally improves reporting quality, though their effectiveness varies and is generally quite low. The impact of the use of the CONSORT statement has been assessed in a Cochrane review including data from more than 16,600 trials[59]. Between endorsing and non-endorsing journals, 25 outcomes out of 27 were improved with endorsement, though only five of these significantly. For (much smaller) analyses of the impact of other reporting guideline introduction, modest improvements have been seen over time, though not attributable to endorsement of the reporting guidelines[44–46]. Active implementation of reporting checklists, such as requiring the completion of reporting checklists with article submissions[199] and editorial use of checklists[200], appears to have resulted in more pronounced improvements.

There are potential issues with existing evaluations of reporting guidelines. Many assessments of reporting quality focus on an overall reporting rate against a checklist, equally weighting each item. If we are interested in transparency as a means to achieve greater reproducibility[f], this approach is not always appropriate. An improvement in score does not necessarily mean that there is an improvement in reproducibility: if just a single key methodological item is not reported, repeating the experiment may not be possible, even if an article is otherwise well reported. The interpretation of such scores can be ambiguous; for example, in one review of compliance with the QUOROM checklist, reporting quality was described as “basically acceptable” with 68% compliance, despite the fact that reporting was “unsatisfactory” in items related to the methods section[200]. Many would draw a different conclusion. If assessing methods reproducibility[201], collapsing overall scores per article into a binary assessment of “reproducible” or “not reproducible” may sometimes be more appropriate. When examining transparency in general, overall scores may be preferred. In Chapter 2, we provided both of these by conducting an analysis of essential methodological items as well as overall scores.

---

[f] We recognise that this is not always the case. In clinical trials for example, improved transparency so that the risk of bias can be better assessed, or so that results can be synthesised into higher level evidence, may make increased transparency an end in itself.

In some cases, again if reproducibility is a focus, assessing reporting quality may be unnecessary and reproducibility could be assessed directly. In computational research, which can be deterministic, the reproducibility of both methods and results can be demonstrated by using the same inputs in the same model and assessing whether or not the results are equivalent[201]. In other fields, such an approach is not feasible because there is noise in the methods and results: for example, even if two RCTs used identical methods, some variation in participant characteristics would lead to at least some variation in results. Practical issues, such as resource requirements, are also apparent.

Overall, the evidence suggests that reporting guidelines can have a positive influence on reporting quality. However, passive endorsement and even active implementation of reporting guidelines may not be sufficient to provide the magnitude of change for which many would hope. Given the large proliferation of reporting guidelines, including ours, further evaluation of their effectiveness is warranted, and research into improvements in implementation would be useful. Where reproducibility is the aim, focussing on deterministic areas of research to establish the effectiveness of reporting guidance could be valuable. Lessons could then be applied to other areas.

6.3 Patents in Literature Searches

Chapters 2 and 3 focussed on patent landscape reporting quality. Here, we discuss the potential utility of analysis of patents for researchers more generally rather than focussing on reporting quality.

Academic literature searches are considered an important part of most scientific research. Research articles include comparisons to or summaries of the literature, there are a huge number of review articles published, and systematic or scoping reviews are increasingly prevalent[202]. Often, the premise is that it is important to understand the existing literature such that new work can build on or differentiate from it. It is surprising, therefore, that the prevalence of reviews or analysis of patents within academia is far less common. More than 20% of scientific researchers in academia, government and non-profit sectors have never read a patent[203].

From an academic perspective, patents are a rich and openly accessible source of information, in some areas having little overlap with the coverage of academic articles[12]. Although there are concerns over the transparency, accuracy and timeliness of publication of patent documents, similar concerns exist for the academic literature[7,31], and it is still considered useful. The cost of patenting is higher than publishing academically, so, in academia, patents may only be filed if there is an expectation of commercial value. Despite this, very few papers cite patents. In commercial research, experiments published in patents are often never published in academic journals, though the extent of this is unknown[9]. When reviewing the literature, simple searches of the patent literature could provide valuable insight.

Patent landscaping enables the summarisation and analysis of multiple patent documents. Setting aside methodological or reporting issues, such analyses could be useful in understanding the state-of-the-art in a particular area. Comparing the patent and academic literature might allow areas of perceived commercial value to be identified, and to identify areas in which intellectual property restrictions render the development of new technologies challenging. Such analyses might be more useful for strategic decisions, such as those made by funders or government, than for individual researchers. Indeed, the UK government has published several patent landscapes[60].

Depending on the aim of a literature review or analysis, therefore, some searching of the patent literature might be warranted. In Chapter 4, we noted that article pre-prints might also represent an under-used source of information for systematic reviews. Useful information sources sometimes include those that have not traditionally been searched, and when developing search strategies for literature reviews it may be worth considering this.

6.4 Predictive Modelling of Regulatory Approval

Some limitations and outstanding work for our models were discussed in Chapter 5. Here, we discuss more general criticisms that may be raised at approaches to predict regulatory approval and suggest areas for future work.

6.4.1 Potential Criticisms

Given the extremely long feedback cycles required to fully evaluate a predictive model of regulatory approval (for example, the average time from preclinical testing to approval is 12 years[204]), it could be argued that its utility is limited. The same argument could be made in the context of the end user of the model: who would want to use a predictive model that cannot be validated for at least ten years? However, the same criticism can be applied to almost any technology used in drug discovery or development, including high throughput screening, animal models, genomics, computational drug design and discovery, and many others, at the point they are first introduced. If we assume that the ultimate aim of drug discovery and development is to achieve regulatory approval (discussed below), then a large amount of technology used in R&D is implicitly predicting approval and is only truly validated when the compounds evaluated using it are approved. Despite this, many technologies rapidly proliferate. Indeed, several criticisms have attributed the decline in R&D efficiency to increasing use of, and enormous investment in, reductionist models without sufficient consideration to their predictive validity in a clinical setting[2,18,19,205]. Our model, by contrast, would cost almost nothing to implement and explicitly predicts approval. Because of this, we already have an estimate as to how it will perform against this metric. We do not need to rely on alternative metrics with unquantified relationships to the outcome of interest.

Though regulatory approval is often considered the “gold standard” outcome for drug development, it is an imperfect outcome measure[2,18]. In reality, drug success is not binary. Revenues, number of patients treated, severity of unmet medical need, competition, adherence, etc. all differ between approved drugs and influence their importance to drug developers or society. If a model consistently prioritised compounds that achieve approval but have low value at the expense of high value drugs with lower approval probability, it could have negative utility. Incorporating these variables into a predictive model, however, would be challenging and probably require data that are not readily available. Instead, they should be considered and possibly modelled in impact evaluations of predictive models of approval. Assessing the probability of approval is just one part of the decision-making process.

On a related note, a possible perceived shortcoming of our approach and others similar is that they cannot predict step changes in innovation. It could be argued that by relying on this model, which will tend to rank drug candidates with parameters more similar to approved drugs more highly, R&D portfolios could become biased against innovative ideas. Such innovative ideas might have a low probability of success but huge potential value if successful. However, if we accept that drug candidates more similar to approved drugs have a higher probability of becoming approved drugs (which is the premise of empirical approaches to predictive modelling in this space), then this criticism is easily addressed. Our model can assist in determining the probability of approval, which will surely be lower for highly innovative approaches, and the user can take into account other considerations either qualitatively or quantitatively. It may be that a portfolio of innovative compounds has a higher expected utility than a portfolio of “drug-like” compounds, but it would still not be reasonable to suggest that the innovative approaches had a higher probability of approval.

Some may argue that the process of drug development is too complex to be modelled meaningfully. This may be true; any model based on the predictor variables we have used will be a dramatic over-simplification of reality. However, the alternative is to rely on expert opinion. In medicinal chemistry, the lack of consensus among experts over “chemically attractive” compounds or fragments is very high[117]. For example, in one analysis, 19 chemists were asked to select or reject 4,000 compounds. Consensus, defined as 75% agreement, was reached on just 8% of compounds: 7% of those rejected and 1% selected[206]. It seems likely that looking at non-chemical predictors, such as those included in our discovery model or those in other models[125,127], the same would be true. As should be clear from the previous paragraphs, we do not suggest relying solely on the output of predictive models of approval. However, computational models can at least apply rules and calculations consistently[117]. Determining those rules and calculations based on analysis of approved and discontinued drugs is logical.

6.4.2 Avenues for Future Work

Assuming that the premise of the approach is accepted, there are several areas where the model might be improved. These fall into two categories: different modelling approaches and changes to predictor variables.

Logistic regression models were developed from our datasets. Logistic regression is a commonly used and recommended approach for binary classification problems[95,207], but there are many other, increasingly commonly used algorithms available[177]. Random forests is a particularly common approach used in all three papers identified in Chapter 4. Lo *et al.* directly compare the performance of penalised logistic regression to support vector machine and random forests, and find random forests to give the best performance[127]. Heinemann *et al.* also compare several algorithms, including random forests but not logistic regression, and find random forests performs best[126]. A large benchmarking study of logistic regression in comparison to random forests showed that random forests performed marginally better on average[208]. However, there is also evidence suggesting that unless very large datasets are available (> 200 events per variable [EPV]), logistic regression gives similar and less optimistic performance over common machine learning algorithms, including random forests[128]. In all cases, improvements with other modelling approaches (if any) are likely to be marginal. Further, machine learning approaches in drug-likeness have been criticised for lacking intuitiveness, transparency and ease of implementation[145]. Nevertheless, exploring them may be worthwhile because even small improvements in predictive performance could have a significant impact on the overall costs of drug development[18].

We do not account for time in our analysis and modelling approaches that do might be useful because drug characteristics may change over time. For example, there is very recent evidence of small changes for some physicochemical characteristics of approved drugs over time[189]. To address this issue, Lo *et al.* developed models in five-year time windows, using only data from that time period for development and documenting changes in performance over time[127]. However, given that EPV for their whole dataset is (optimistically) between four and five (when ten is the minimum recommended), these sub-analyses will have very low EPV and therefore are likely to be overfit. The same issue prevents us from performing such an analysis with the discovery dataset. With the physicochemical dataset, we cannot do this analysis because we do not have the start dates for all drugs to generate the time windows. If larger datasets with the required information can be generated, analyses accounting for time would be valuable.

The predictors used in our model are relatively simplistic and with increased resources more descriptive variables would be feasible. Our literature count variables only capture a small proportion of the available information. Text mining approaches[9,209] could potentially be used to generate more informative predictors, for example, relating to the authorship (e.g. academic vs. commercial), funding, or conclusions (e.g. significant effect vs. not). Additionally, information regarding indications and targets could be included. Success rates differ across indications and regulatory designations[111], so including this information in the model could result in better calibration and greater accuracy. Drug target information can be used to identify problematic drugs[135] and again might be informative. This list is not exhaustive and, in each case, a large number of predictor variables could be generated.

The problem, therefore, is that without a considerably larger dataset (these predictors are mainly relevant to the discovery dataset, n = 1,220, 212 approved), EPV constraints will be quickly violated. The primary limitation of adding predictors to the discovery model will be gathering a larger dataset with a development history that allows determination of the dates before which predictor variable data should be collected. Additional commercial databases may be useful for this (e.g. that used in ref: [127]). Alternatively, a different moment of prediction could be used with our data (e.g. commencement of phase II trials), which would increase the number of observations with appropriate data.

Additional physicochemical predictors could be added to the physicochemical dataset. As mentioned in Chapter 5, we are working to add the number of structural alerts, which would allow a comparison to a well-known model of drug-likeness[145] to be performed. A large range of other properties can be readily calculated and have been used in other models[181]. Because this dataset is larger and currently has few predictor variables, adding variables should be less problematic and might represent a promising next step.

In developing our models, we considered only predictors that would be relevant prior to the commencement of any clinical studies. If, however, a moment of prediction later in development was used, different and potentially more informative predictors could be used, for example relating to the characteristics of trials, regulatory designation, and competition for the indication or target. Such information would likely improve the performance of the model in terms of AUC, but would need to be weighed against the potential cost of making predictions later in development. Our systematic review

(Chapter 4) did not identify any models suitable for predicting approval from the commencement or end of phase I clinical trials, so future models may wish to explore this moment.

6.5 Concluding Remarks

This thesis began by outlining the evidence that the efficiency of biomedical research has been consistently falling. Much of the evidence we have accumulated provides further explanation for potential causes of this inefficiency: i) availability of suitable data, ii) the quality of analysis of that data, and iii) the quality of reporting of that analysis.

i. The ability to conduct useful analysis relies on the availability of data that allows us to conduct it. Although one of the goals of the patent system is to disseminate scientific knowledge, US patents have only been made available for bulk download by the US Patent and Trade Mark Office since 2010. Other avenues for searching and analysing patent data have existed for some time, though the most advanced capabilities require subscriptions and are therefore not widely accessible. Two[125,127] of the three papers included in Chapter 4 highlighted the need for more or better quality data to generate better models. In the development of our model in Chapter 5, data availability was also a challenge. Though increasingly comprehensive open databases exist for approved and investigational drugs[210], as well as withdrawn drugs[211], collated data on compounds that are discontinued during development is not openly available. Data on individual discontinued drugs is available publicly and collating it could be useful to allow other research to compare approved and failed drugs. In general, continued efforts to make useful data freely available are essential.

ii. Across a number of fields touched on in this thesis, analyses conducted can be improved or made more efficient. We noted in Chapter 2 that many of the patent landscapes were relatively simplistic and those employing more complex approaches may not appropriately answer the intended questions. Given that patent information is readily available, a focus in this area on improving methodologies and on conducting more hypothesis-driven research might be valuable. In predicting drug approvals, several methodological issues were consistently identified. In clinical prediction modelling, these issues have been widely reported and guidance developed to address them[95]. If we recognise that the appropriate types of analysis are essentially the same between applications, these guidelines could be used much more widely. To give a concrete example, issues with binning continuous variables in the specific context of drug-likeness became widely recognised in 2013 (ref: [138]); essentially the same issues have been recognised in the context of clinical prognostic modelling since at least 1994 (ref: [212]). The same may be true of analysis outside prediction modelling.

iii. Regardless of the available data and quality of analysis, the published report is the primary means by which research can be assessed, synthesised, and acted on. This thesis finds evidence that the widely reported deficiencies in reporting of health research are present more broadly in the life sciences. Because the raw data in patent landscape articles (i.e. patent documents) are freely available, there is little reason not to have full transparency and

> reproducibility in published articles, for example by including the full list of patents analysed. In modelling regulatory approval, presentation of the raw data may not be feasible, as databases used can be proprietary (ref: [127] and our model). In these cases, clearly reporting the model in a way that allows its usage is essential so that others can attempt to externally validate it. Without the final model or the underlying data, evaluating the model is challenging. As we have discussed, the solution to reporting quality is not clear, and more research into evaluating and implementing guidance (or other solutions) is needed. As with quality of analysis, this research may first be conducted in the context of health science where the issues are well characterised. As evidence accumulates, encouraging cross-disciplinary applications of the findings will again be useful.

We have taken best-practices from clinical and health research and attempted to apply them to broad issues that may impact biomedical research, specifically by: i) developing a reporting guideline to improve reporting quality of patent landscapes and ii) developing a transparently reported predictive model of regulatory approval that may help to prioritise compounds. Regardless of the outstanding work, this model can be externally validated based on the data presented. In this discussion, we have made several additional suggestions that might help to improve the quality and efficiency of research being conducted. Several areas for future work have also been outlined.

7 References

1. Bowen, A. & Casadevall, A. Increasing disparities between resource inputs and outcomes, as measured by certain health deliverables, in biomedical research. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11335–11340 (2015).
2. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191–200 (2012).
3. DiMasi, J. A., Grabowski, H. G. & Hansen, R. W. Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33 (2016).
4. Thakor, R. T. *et al.* Just how good an investment is the biopharmaceutical sector? *Nat. Biotechnol.* **35**, 1149–1157 (2017).
5. Berndt, E. R., Nass, D., Kleinrock, M. & Aitken, M. Decline in economic returns from new drugs raises questions about sustaining innovations. *Health Aff. (Millwood)* **34**, 245–252 (2015).
6. Baker, M. Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the ‘crisis’ rocking science and what they think will help. *Nature* **533**, 452–455 (2016).
7. Ioannidis, J. P. A. Why most published research findings are false. *PLOS Med.* **2**, e124 (2005).
8. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* **2018**, (2018).

9. Rodriguez-Esteban, R. & Bundschus, M. Text mining patents for biomedical knowledge. *Drug Discov. Today* **21**, 997–1002 (2016).
10. World Intellectual Property Indicators. (2015).
11. Plume, A. & van Weijen, D. Publish or perish? The rise of the fractional author…. *Res. Trends* **38**, (2014).
12. Southan, C., Várkonyi, P. & Muresan, S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. *J. Cheminformatics* **1**, 10 (2009).
13. Reardon, S. Text-mining offers clues to success. *Nat. News* **509**, 410 (2014).
14. Schneider, N., Lowe, D. M., Sayle, R. A., Tarselli, M. A. & Landrum, G. A. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *J. Med. Chem.* **59**, 4385–4402 (2016).
15. Roberts, M. *et al.* The global intellectual property landscape of induced pluripotent stem cell technologies. *Nat. Biotechnol.* **32**, 742–748 (2014).
16. Clark, K. *et al.* Patent data mining: A tool for accelerating HIV vaccine innovation. *Vaccine* **29**, 4086–4093 (2011).
17. Bubela, T. *et al.* Patent landscaping for life sciences innovation: toward consistent and transparent practices. *Nat. Biotechnol.* **31**, 202–206 (2013).
18. Scannell, J. W. & Bosley, J. When quality beats quantity: decision theory, drug discovery, and the reproducibility crisis. *PLOS One* **11**, e0147215 (2016).
19. Horrobin, D. F. Modern biomedical research: an internally self-consistent universe with little contact with medical reality? *Nat. Rev. Drug Discov.* **2**, 151 (2003).

20. Smith, J. A., Arshad, Z., Thomas, H., Carr, A. J. & Brindley, D. A. Evidence of insufficient quality of reporting in patent landscapes in the life sciences. *Nat. Biotechnol.* **35**, 210–214 (2017).
21. Smith, J. *et al.* Performance and quality of algorithms for prediction of therapeutic market authorisation: systematic review protocol. CRD42018093735. *PROSPERO* (2018).
22. Johnson, T. S. *et al.* Genetic improvement of biofuel plants: recent progress and patents. *Recent Pat. DNA Gene Seq.* **7**, 2–12 (2013).
23. Panja, S., Majumder P, P., Sarkar, B. K., Mukim, K. K. & Hati, A. Global research on medical cotton - Evidence from patent landscape study. *J. Intellect. Prop. Rights* **20**, 39–50 (2015).
24. Fiala, J. L. A. & Lowery, D. Patent watch: migraine therapies targeting the CGRP pathway: intellectual property landscape. *Nat. Rev. Drug Discov.* **15**, 8–9 (2016).
25. Egelie, K. J., Graff, G. D., Strand, S. P. & Johansen, B. The emerging patent landscape of CRISPR-Cas gene editing technology. *Nat. Biotechnol.* **34**, 1025–1031 (2016).
26. Paradise, J., Andrews, L. & Holbrook, T. Intellectual property. Patents on human genes: an analysis of scope and claims. *Science* **307**, 1566–1567 (2005).
27. Jaenichen, H.-R. & Pitz, J. Research exemption/experimental use in the European Union: patents do not block the progress of science. *Cold Spring Harb. Perspect. Med.* **5**, a020941 (2014).
28. Rai, A. K. & Sherkow, J. S. The changing life science patent landscape. *Nat. Biotechnol.* **34**, 292–294 (2016).
29. Edwards, A. Perspective: Science is still too closed. *Nature* **533**, S70–S70 (2016).

30. Schulz, K. F., Altman, D. G. & Moher, D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ* **340**, c332 (2010).

31. Altman, D. G. & Moher, D. Importance of transparent reporting of health research. in *Guidelines for Reporting Health Research: A User's Manual* (eds. Moher, D., Altman, D. G., Schulz, K. F., Simera, I. & Wager, E.) 1–13 (John Wiley & Sons, Ltd, 2014).

32. von Elm, E. *et al.* The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *The Lancet* **370**, 1453–1457 (2007).

33. Liberati, A. *et al.* The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *BMJ* **339**, b2700 (2009).

34. Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ* **339**, b2535 (2009).

35. Grant, E., Van den Hof, M. & Gold, E. R. Patent landscape analysis: a methodology in need of harmonized standards of disclosure. *World Pat. Inf.* **39**, 3–10 (2014).

36. Gold, E. R. & Baker, A. M. Evidence-based policy: understanding the technology landscape. *J. Law Inf. Sci.* **22**, (2012).

37. Definition of life sciences in English. *Oxford Dictionaries* Available at: https://en.oxforddictionaries.com/definition/life_sciences. (Accessed: 1st November 2016)

38. Jordan, K. P. & Lewis, M. Improving the quality of reporting of research studies. *Musculoskeletal Care* **7**, 137–142 (2009).

39. Chalmers, I. & Glasziou, P. Avoidable waste in the production and reporting of research evidence. *The Lancet* **374**, 86–89 (2009).

40. Schulz, K. F., Chalmers, I., Hayes, R. J. & Altman, D. G. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA* **273**, 408–412 (1995).

41. Garfield, E. Journal impact factor: a brief review. *Can. Med. Assoc. J.* **161**, 979–980 (1999).

42. Devereaux, P. J., Manns, B. J., Ghali, W. A., Quan, H. & Guyatt, G. H. The reporting of methodological factors in randomized controlled trials and the association with a journal policy to promote adherence to the Consolidated Standards of Reporting Trials (CONSORT) checklist. *Control. Clin. Trials* **23**, 380–388 (2002).

43. Montané, E., Vallano, A., Vidal, X., Aguilera, C. & Laporte, J.-R. Reporting randomised clinical trials of analgesics after traumatic or orthopaedic surgery is inadequate: a systematic review. *BMC Clin. Pharmacol.* **10**, 2 (2010).

44. Prady, S. L., Richmond, S. J., Morton, V. M. & Macpherson, H. A systematic evaluation of the impact of STRICTA and CONSORT recommendations on quality of reporting for acupuncture trials. *PLOS One* **3**, e1577 (2008).

45. Smidt, N. *et al.* The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology* **67**, 792–797 (2006).

46. Leung, V., Rousseau-Blass, F., Beauchamp, G. & Pang, D. S. J. ARRIVE has not ARRIVEd: support for the ARRIVE (animal research: reporting of in vivo experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. *PLOS One* **13**, e0197882 (2018).

47. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348**, 1422–1425 (2015).

48. McNutt, M. Journals unite for reproducibility. *Science* **346**, 679–679 (2014).

49. Adelman, D. E. & DeAngelis, K. L. Patent metrics: the mismeasure of innovation in the biotech patent debate. *Tex. Law Rev.* **85**, 1–79 (2006).

50. Arshad, Z. *et al.* Open access could transform drug discovery: a case study of JQ1. *Expert Opin. Drug Discov.* **11**, 321–332 (2016).

51. Kalpana Sastry, S., Rashmi, H. B. & Badri, J. Research and development perspectives of transgenic cotton: Evidence from patent landscape studies. *J. Intellect. Prop. Rights* **16**, 139–153 (2011).

52. Cucoranu, I. C., Parwani, A. V., Vepa, S., Weinstein, R. S. & Pantanowitz, L. Digital pathology: a systematic evaluation of the patent landscape. *J. Pathol. Inform.* **5**, 16 (2014).

53. Swamy, H. M. M. *et al.* Analysis of opportunities and challenges in patenting of Bacillus thuringiensis insecticidal crystal protein genes. *Recent Pat. DNA Gene Seq.* **6**, 64–71 (2012).

54. Glasziou, P., Meats, E., Heneghan, C. & Shepperd, S. What is missing from descriptions of treatment in trials and reviews? *BMJ* **336**, 1472–1474 (2008).

55. Reveiz, L. *et al.* Reporting of methodologic information on trial registries for quality assessment: a study of trial records retrieved from the WHO search portal. *PLOS One* **5**, (2010).

56. Fleming, P. S., Koletsi, D., Polychronopoulou, A., Eliades, T. & Pandis, N. Are clustering effects accounted for in statistical analysis in leading dental specialty journals? *J. Dent.* **41**, 265–270 (2013).

57. Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C. & Altman, D. G. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA* **291**, 2457–2465 (2004).
58. Kesselheim, A. S., Wang, B., Studdert, D. M. & Avorn, J. Conflict of interest reporting by authors involved in promotion of off-label drug use: an analysis of journal disclosures. *PLOS Med.* **9**, e1001280 (2012).
59. Turner, L. *et al.* Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database Syst. Rev.* **11**, MR000030 (2012).
60. UK Intellectual Property Office. Eight great technologies: the patent landscapes. (2014).
61. World Intellectual Property Office (WIPO). Patent landscape report on microalgae-related technologies. (2016).
62. UK Intellectual Property Office. Graphene: the worldwide patent landscape in 2015. (2015).
63. UK Intellectual Property Office. Eight great technologies: regenerative medicine. (2014).
64. Sternitzke, C. An exploratory analysis of patent fencing in pharmaceuticals: The case of PDE5 inhibitors. *Res. Policy* **42**, 542–551 (2013).
65. Bergman, K. & Graff, G. D. The global stem cell patent landscape: implications for efficient technology transfer and commercial development. *Nat. Biotechnol.* **25**, 419–424 (2007).
66. Jensen, K. & Murray, F. Intellectual property landscape of the human genome. *Science* **310**, 239–240 (2005).

67. Moher, D., Tetzlaff, J., Tricco, A. C., Sampson, M. & Altman, D. G. Epidemiology and reporting characteristics of systematic reviews. *PLOS Med.* **4**, e78 (2007).
68. Chan, A.-W. & Altman, D. G. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* **365**, 1159–1162 (2005).
69. Neumann, P. J., Stone, P. W., Chapman, R. H., Sandberg, E. A. & Bell, C. M. The quality of reporting in published cost-utility analyses, 1976-1997. *Ann. Intern. Med.* **132**, 964–972 (2000).
70. Kilkenny, C. *et al.* Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLOS One* **4**, e7824 (2009).
71. Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M. & Altman, D. G. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLOS Biol.* **8**, e1000412 (2010).
72. Husereau, D. *et al.* Consolidated health economic evaluation reporting standards (CHEERS) statement. *Eur. J. Health Econ.* **14**, 367–372 (2013).
73. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
74. Goodacre, R. *et al.* Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **3**, 231–241 (2007).
75. Plint, A. C. *et al.* Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med. J. Aust.* **185**, (2006).
76. Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O. & Alberti, C. Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review. *PLOS One* **6**, e20476 (2011).

77. Hasson, F., Keeney, S. & McKenna, H. Research guidelines for the Delphi survey technique. *J. Adv. Nurs.* **32**, 1008–1015 (2000).

78. Powell, C. The Delphi technique: myths and realities. *J. Adv. Nurs.* **41**, 376–382 (2003).

79. Giannarou, L. & Zervas, E. Using Delphi technique to build consensus in practice. *Int. J. Bus. Sci. Appl. Manag.* **9**, 65–82 (2014).

80. Simera, I., Altman, D. G., Moher, D., Schulz, K. F. & Hoey, J. Guidelines for reporting health research: the EQUATOR network's survey of guideline authors. *PLOS Med.* **5**, e139 (2008).

81. Bolarinwa, O. A. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. *Niger. Postgrad. Med. J.* **22**, 195–201 (2015).

82. Burns, K. E. A. & Kho, M. E. How to assess a survey report: a guide for readers and peer reviewers. *CMAJ Can. Med. Assoc. J.* **187**, E198–E205 (2015).

83. Heiko, A. Consensus measurement in Delphi studies: review and implications for future quality assurance. *Technol. Forecast. Soc. Change* **79**, 1525–1536 (2012).

84. Scheibe, M., Skutsch, M. & Schofer, J. Experiments in delphi methodology. in *The Delphi Method: Techniques and Applications* **18**, (1975).

85. Viera, A. J. & Garrett, J. M. Understanding interobserver agreement: the kappa statistic. *Fam. Med.* **37**, 360–363 (2005).

86. Moher, D. *et al.* Describing reporting guidelines for health research: a systematic review. *J. Clin. Epidemiol.* **64**, 718–742 (2011).

87. Bowman, P. A. & Greenleaf, D. Non-CFC metered dose inhalers: the patent landscape. *Int. J. Pharm.* **186**, 91–94 (1999).

88. Lundin, P. Is silence still golden? Mapping the RNAi patent landscape. *Nat. Biotechnol.* **29**, 493–497 (2011).

89. Petering, J., McManamny, P. & Honeyman, J. Antibody therapeutics - the evolving patent landscape. *New Biotechnol.* **28**, 538–544 (2011).

90. Agarwal, P. & Searls, D. B. Can literature analysis identify innovation drivers in drug discovery? *Nat. Rev. Drug Discov.* **8**, 865 (2009).

91. Korting, H. C., Blecher, P., Schäfer-Korting, M. & Wendel, A. Topical liposome drugs to come: what the patent literature tells us. A review. *J. Am. Acad. Dermatol.* **25**, 1068–1071 (1991).

92. World Intellectual Property Office (WIPO). Guidelines for preparing patent landscape reports. (2015).

93. Oldman, P., Kitsara, I. & World Intellectual Property Office (WIPO). The WIPO Manual on Open Source Patent Analytics. (2016).

94. Shamseer, L. *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* **349**, g7647 (2015).

95. Moons, K. G. M. *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* **162**, W1 (2015).

96. Husereau, D. *et al.* Consolidated health economic evaluation reporting standards (CHEERS) - explanation and elaboration: a report of the ISPOR health economic evaluation publication guidelines good reporting practices task force. *Value Health J. Int. Soc. Pharmacoeconomics Outcomes Res.* **16**, 231–250 (2013).

97. Pereira, C. G., Picanco-Castro, V., Covas, D. T. & Porto, G. S. Patent mining and landscaping of emerging recombinant factor VIII through network analysis. *Nat. Biotechnol.* **36**, 585–590 (2018).

98. Abud, M. J., Hall, B. & Helmers, C. An empirical analysis of primary and secondary pharmaceutical patents in Chile. *PLOS One* **10**, e0124257 (2015).

99. Endacott, J. & Poolman, R. Looking for insights – quality control initiatives for enhancing patent searches. *World Pat. Inf.* **1**, 3–7 (2013).

100. Yang, Y. Y., Akers, L., Yang, C. B., Klose, T. & Pavlek, S. Enhancing patent landscape analysis with visualization output. *World Pat. Inf.* **32**, 203–220 (2010).

101. Tseng, Y.-H., Lin, C.-J. & Lin, Y.-I. Text mining techniques for patent analysis. *Inf. Process. Manag.* **43**, 1216–1247 (2007).

102. Drazen, J. M. *et al.* Uniform format for disclosure of competing interests in ICMJE journals. *N. Engl. J. Med.* **361**, 1896–1897 (2009).

103. Bell, C. M. *et al.* Bias in published cost effectiveness studies: systematic review. *BMJ* **332**, 699–703 (2006).

104. Black, N. The Cooksey review of UK health research funding. *BMJ* **333**, 1231 (2006).

105. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for developers of health research reporting guidelines. *PLOS Med.* **7**, (2010).

106. Hsu, C.-C. & Sandford, B. A. The Delphi technique: making sense of consensus. *Pract. Assess. Res. Eval.* **12**, 1–8 (2007).

107. Moher, D. *et al.* Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst. Rev.* **4**, 1 (2015).

108. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191–200 (2012).

109. Hay, M., Thomas, D. W., Craighead, J. L., Economides, C. & Rosenthal, J. Clinical development success rates for investigational drugs. *Nat. Biotechnol.* **32**, 40–51 (2014).

110. Thomas, D. *et al. Clinical Development Success Rates 2006-2015*. (San Diego: Biomedtracker, 2016).

111. Wong, C. H., Siah, K. W. & Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics* kxx069 (2018).

112. Denayer, T., Stöhr, T. & Van Roy, M. Animal models in translational medicine: validation and prediction. *New Horiz. Transl. Med.* **2**, 5–11 (2014).

113. Shanks, N., Greek, R. & Greek, J. Are animal models predictive for humans? *Philos. Ethics Humanit. Med. PEHM* **4**, 2 (2009).

114. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).

115. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **1**, 337–341 (2004).

116. Hung, C.-L. & Chen, C.-C. Computational approaches for drug discovery. *Drug Dev. Res.* **75**, 412–418 (2014).

117. Cumming, J. G., Davis, A. M., Muresan, S., Haeberlein, M. & Chen, H. Chemical predictive modelling to improve compound quality. *Nat. Rev. Drug Discov.* **12**, 948–962 (2013).

118. Ursu, O., Rayan, A., Goldblum, A. & Oprea, T. I. Understanding drug-likeness. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 760–781 (2011).

119. Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **23**, 302–321 (2003).

120. Higgins, J. P. & Green, S. *Cochrane handbook for systematic reviews of interventions*. **4**, (John Wiley & Sons, 2011).

121. Reviews, U. of Y. C. for & Dissemination. *Systematic reviews: CRD's guidance for undertaking reviews in health care*. (University of York, Centre for Reviews & Dissemination, 2009).

122. Moons, K. G. M. *et al.* Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLOS Med.* **11**, e1001744 (2014).

123. Wartolowska, K. A. *et al.* The magnitude and temporal changes of response in the placebo arm of surgical randomized controlled trials: a systematic review and meta-analysis. *Trials* **17**, 589 (2016).

124. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.* **162**, 55–63 (2015).

125. DiMasi, J. A. *et al.* A tool for predicting regulatory approval after phase II testing of new oncology compounds. *Clin. Pharmacol. Ther.* **98**, 506–513 (2015).

126. Heinemann, F., Huber, T., Meisel, C., Bundschus, M. & Leser, U. Reflection of successful anticancer drug development processes in the literature. *Drug Discov. Today* **21**, 1740–1744 (2016).

127. Lo, A. W., Siah, K. W. & Wong, C. H. Machine-learning models for predicting drug approvals and clinical-phase transitions. *SSRN* (2017).

128. van der Ploeg, T., Austin, P. C. & Steyerberg, E. W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med. Res. Methodol.* **14**, 137 (2014).

129. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Int. Jt. Conf. Articial Intell.* **14**, 1137–1145 (1995).

130. Steyerberg, E. W. & Jr, F. E. H. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* 8 (2001).

131. Malik, L. *et al.* Predicting success in regulatory approval from Phase I results. *Cancer Chemother. Pharmacol.* **74**, 1099–1103 (2014).

132. Goffin, J., Baral, S., Tu, D., Nomikos, D. & Seymour, L. Objective responses in patients with malignant melanoma or renal cell cancer in early clinical studies do not predict regulatory approval. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **11**, 5928–5934 (2005).

133. El-Maraghi, R. H. & Eisenhauer, E. A. Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **26**, 1346–1354 (2008).

134. Jardim, D. L., Groves, E. S., Breitfeld, P. P. & Kurzrock, R. Factors associated with failure of oncology drugs in late-stage clinical development: A systematic review. *Cancer Treat. Rev.* **52**, 12–21 (2017).

135. Lopes, T. J. S., Shoemaker, J. E., Matsuoka, Y., Kawaoka, Y. & Kitano, H. Identifying problematic drugs based on the characteristics of their targets. *Front. Pharmacol.* **6**, 186 (2015).

136. Schachter, D. Master's Thesis: Probabilisitic modeling of the drug development domain: a bayesian domain-knowledge application for pharmacovigilance. (Massachusetts Institute of Technology, 2003).

137. Lopes, T. J. D. S., Kitano, H. & Kawaoka, Y. US20150242752A1: Approval prediction apparatus, approval prediction method, and computer program product. (2015).

138. Kenny, P. W. & Montanari, C. A. Inflation of correlation in the pursuit of drug-likeness. *J. Comput. Aided Mol. Des.* **27**, 1–13 (2013).

139. Lampe, J. & Konieczny, A. US20110196620A1: Opportunity sector analysis tool. (2011).

140. Segall, M. & Hashimoto, T. US9367812B2: Compound selection in drug discovery. (2016).

141. Iwema, C. L., LaDue, J., Zack, A. & Chattopadhyay, A. search.bioPreprint: a discovery tool for cutting edge, preprint biomedical research articles. *F1000Research* **5**, (2016).

142. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **9**, 203–214 (2010).

143. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25 (1997).

144. Lipinski, C. A. Rule of five in 2015 and beyond: Target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions. *Adv. Drug Deliv. Rev.* **101**, 34–41 (2016).

145. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
146. Ursu, O. & Oprea, T. I. Model-free drug-likeness from fragments. *J. Chem. Inf. Model.* **50**, 1387–1394 (2010).
147. Tang, K., Zhu, R., Li, Y. & Cao, Z. Discrimination of approved drugs from experimental drugs by learning methods. *BMC Bioinformatics* **12**, 157 (2011).
148. R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. (2018).
149. Wickham, H. *Tidyverse: easily install and load the 'Tidyverse'. R package version 1.2.1*. (2017).
150. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **45**, 2615–2623 (2002).
151. Kim, S. *et al.* PubChem substance and compound databases. *Nucleic Acids Res.* **44**, D1202–D1213 (2016).
152. Cheng, T. *et al.* Computation of octanol-water partition coefficients by guiding an additive model with knowledge. *J. Chem. Inf. Model.* **47**, 2140–2148 (2007).
153. Wang, R., Fu, Y. & Lai, L. A new atom-additive method for calculating partition coefficients. *J. Chem. Inf. Comput. Sci.* **37**, 615–621 (1997).
154. World Health Organization (WHO). International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10). (2016).
155. Windsor, D. A. Could bibliometric data be used to predict the clinical success of drugs? *J. Doc.* **32**, 174–181 (1976).
156. Kissin, I. What can big data on academic interest reveal about a drug? Reflections in three major us databases. *Trends Pharmacol. Sci.* **39**, 248–257 (2018).

157. UK Intellectual Property Office. Patent fact sheets. (2014).

158. Eurostat. Venture capital investments in HTEC sectors. (2016). Available at: http://ec.europa.eu/eurostat/web/science-technology-innovation/data/database. (Accessed: 21st May 2018)

159. Eurostat. Harmonised index of consumer prices. (2018). Available at: http://ec.europa.eu/eurostat/web/hicp/data/database. (Accessed: 21st May 2018)

160. Inflation - up to date info on current and historic inflation by country. (2018). Available at: http://www.inflation.eu/. (Accessed: 21st May 2018)

161. PhRMA. PhRMA Annual Membership Survey. (2017).

162. Global pharmaceutical R&D spending 2008-2022. *Statista* Available at: https://www.statista.com/statistics/309466/global-r-and-d-expenditure-for-pharmaceuticals/. (Accessed: 23rd May 2018)

163. 1-Year Treasury Constant Maturity Rate. (2018). Available at: https://fred.stlouisfed.org/series/DGS1. (Accessed: 22nd May 2018)

164. Bank of England. Bank Rate. (2018). Available at: https://www.bankofengland.co.uk/boeapps/database/Bank-Rate.asp. (Accessed: 22nd May 2018)

165. Peduzzi, P., Concato, J., Feinstein, A. R. & Holford, T. R. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J. Clin. Epidemiol.* **48**, 1503–1510 (1995).

166. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. & Feinstein, A. R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **49**, 1373–1379 (1996).

167. Wynants, L., Collins, G. S. & Calster, B. V. Key steps and common pitfalls in developing and validating risk models. *BJOG Int. J. Obstet. Gynaecol.* **124**, 423–432 (2016).

168. Midi, H., Sarkar, S. K. & Rana, S. Collinearity diagnostics of binary logistic regression model. *J. Interdiscip. Math.* **13**, 253–267 (2010).

169. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).

170. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinforma. Oxf. Engl.* **21**, 3940–3941 (2005).

171. Steyerberg, E. W. *et al.* Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J. Clin. Epidemiol.* **54**, 774–781 (2001).

172. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).

173. Harrell, F. E. *rms: Regression modeling strategies. R package version 5.1-2.* (2018).

174. Liu, A. & Ziebart, B. Robust classification under sample selection bias. in *Advances in neural information processing systems* 37–45 (2014).

175. Zadrozny, B. Learning and evaluating classifiers under sample selection bias. in *In International Conference on Machine Learning ICML'04* 903–910 (2004).

176. Oommen, T., Baise, L. G. & Vogel, R. M. Sampling bias and class imbalance in maximum-likelihood logistic regression. *Math. Geosci.* **43**, 99–120 (2011).

177. Kuhn, M. & Johnson, K. *Applied predictive modeling*. **26**, (Springer, 2013).

178. Altman, N. & Krzywinski, M. Analyzing outliers: influential or nuisance? Points of significance. *Nat. Methods* **13**, 281–282 (2016).

179. Zhang, Z. Residuals and regression diagnostics: focusing on logistic regression. *Ann. Transl. Med.* **4**, (2016).

180. Pajouhesh, H. & Lenz, G. R. Medicinal chemical properties of successful central nervous system drugs. *NeuroRX* **2**, 541–553 (2005).

181. García-Sosa, A. T., Oja, M., Hetényi, C. & Maran, U. DrugLogit: logistic discrimination between drugs and nondrugs including disease-specificity by assigning probabilities based on molecular properties. *J. Chem. Inf. Model.* **52**, 2165–2180 (2012).

182. Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **51**, 817–834 (2008).

183. Congreve, M., Carr, R., Murray, C. & Jhoti, H. A 'rule of three' for fragment-based lead discovery? *Drug Discov. Today* **8**, 876–877 (2003).

184. Hughes, J. D. *et al.* Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* **18**, 4872–4875 (2008).

185. Muthas, D., Boyer, S. & Hasselgren, C. A critical assessment of modeling safety-related drug attrition. *MedChemComm* **4**, 1058–1065 (2013).

186. Ohno, K., Nagahara, Y., Tsunoyama, K. & Orita, M. Are there differences between launched drugs, clinical candidates, and commercially available compounds? *J. Chem. Inf. Model.* **50**, 815–821 (2010).

187. Schachter, A. D., Ramoni, M. F., Baio, G., Roberts, T. G. & Finkelstein, S. N. Economic evaluation of a bayesian model to predict late-phase success of new chemical entities. *Value Health* **10**, 377–385 (2007).

188. Toll, D. B., Janssen, K. J. M., Vergouwe, Y. & Moons, K. G. M. Validation, updating and impact of clinical prediction rules: a review. *J. Clin. Epidemiol.* **61**, 1085–1094 (2008).
189. Shultz, M. D. Two decades under the influence of the rule of five and the changing properties of approved oral drugs. *J. Med. Chem.* Epub ahead of print (2018).
190. DiMasi, J. Risks in new drug development: approval success rates for investigational drugs. *Clin. Pharmacol. Ther.* **69**, 297–307 (2001).
191. Hutton, B. *et al.* The quality of reporting methods and results in network meta-analyses: an overview of reviews and suggestions for improvement. *PLOS One* **9**, e92508 (2014).
192. Jokstad, A., Esposito, M., Coulthard, P. & Worthington, H. V. The reporting of randomized controlled trials in prosthodontics. *Int. J. Prosthodont.* **15**, 230–242 (2002).
193. Dumbrigue, H. B., Jones, J. S. & Esquivel, J. F. Control of bias in randomized controlled trials published in prosthodontic journals. *J. Prosthet. Dent.* **86**, 592–596 (2001).
194. Montenegro, R., Needleman, I., Moles, D. & Tonetti, M. Quality of RCTs in periodontology--a systematic review. *J. Dent. Res.* **81**, 866–870 (2002).
195. Harrison, J. E. Clinical trials in orthodontics II: assessment of the quality of reporting of clinical trials published in three orthodontic journals between 1989 and 1998. *J. Orthod.* **30**, 309–315 (2003).
196. Wang, X. *et al.* Methodology and reporting quality of reporting guidelines: systematic review. *BMC Med. Res. Methodol.* **15**, 15–74 (2015).

197. Chalmers, I., Hedges, L. V. & Cooper, H. A brief history of research synthesis. *Eval. Health Prof.* **25**, 12–37 (2002).

198. Whittle, R., Peat, G., Belcher, J., Collins, G. S. & Riley, R. D. Measurement error and timing of predictor values for multivariable risk prediction models are poorly reported. *J. Clin. Epidemiol.* **102**, 38–49 (2018).

199. Macleod, M. R. & The NPQIP Collaborative group. 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. *bioRvix* (2017).

200. Pandis, N., Shamseer, L., Kokich, V. G., Fleming, P. S. & Moher, D. Active implementation strategy of CONSORT adherence by a dental specialty journal improved randomized clinical trial reporting. *J. Clin. Epidemiol.* **67**, 1044–1048 (2014).

201. Goodman, S. N., Fanelli, D. & Ioannidis, J. P. A. What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12-341ps12 (2016).

202. Ioannidis, J. P. A. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *Milbank Q.* **94**, 485–514 (2016).

203. Ouellette, L. L. Who reads patents? *Nat. Biotechnol.* **35**, 421–424 (2017).

204. Van Norman, G. A. Drugs, devices, and the FDA: Part 1: an overview of approval processes for drugs. *JACC Basic Transl. Sci.* **1**, 170–179 (2016).

205. Horrobin, D. F. Innovation in the pharmaceutical industry. *J. R. Soc. Med.* **93**, 341–345 (2000).

206. Kutchukian, P. S. *et al.* Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. *PLOS One* **7**, e48476 (2012).

207. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: a methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).

208. Couronné, R., Probst, P. & Boulesteix, A.-L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* **19**, 270 (2018).

209. Agarwal, P. & Searls, D. B. Literature mining in support of drug discovery. *Brief. Bioinform.* **9**, 479–492 (2008).

210. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).

211. Siramshetty, V. B. *et al.* WITHDRAWN - a resource for withdrawn and discontinued drugs. *Nucleic Acids Res.* **44**, D1080-1086 (2016).

212. Altman, D. G., Lausen, B., Sauerbrei, W. & Schumacher, M. Dangers of using “optimal” cutpoints in the evaluation of prognostic factors. *JNCI J. Natl. Cancer Inst.* **86**, 829–835 (1994).

8 Supplementary Material

8.1 Chapter 2

Supplementary Table 8.1: Databases searched and search terms used

Database Searched	Search Terms
PubMed	((("Intellectual Property"[Mesh]) OR "Patents as Topic"[Mesh]) AND ((“landscape”) OR (“Analysis”) OR (“Data Mining”)) NOT ((“Foramen”) OR (“Ovale”) OR (“Ductus”) OR (“Arteriosus”))) OR ("Patent") AND ("Landscape")
OvidSP (Embase and Medline)	(Patent OR intellectual property.mp.) AND (methodology OR landscape.mp. OR (data mining.mp. OR data mining) OR (analysis OR data analysis OR content analysis)
Scopus	TITLE-ABS-KEY (("Intellectual Property" OR "Patent") AND ("Landscape" OR "Data Mining")) TITLE (("Intellectual Property" OR " Patent") AND ("Review" OR "Analysis"))
Science Direct	TITLE ("patent" OR "Intellectual Property") AND TITLE – ABSTRACT – KEYWORDS ("Landscape" OR "Analysis" OR "Review" OR "Data Mining")

Supplementary Table 8.2: List of full text papers included in analysis

Abud, M. J., Hall, B. & Helmers, C. An empirical analysis of primary and secondary pharmaceutical patents in Chile. *PLOS One* **10,** e0124257 (2015).
Agarwal, A., Sayres, L. C., Cho, M. K., Cook-Deegan, R. & Chandrasekharan, S. Commercial landscape of noninvasive prenatal testing in the United States: Commercialization of noninvasive prenatal testing. *Prenat. Diagn.* **33,** 521–531 (2013).

Agarwal, P. & Searls, D. B. Can literature analysis identify innovation drivers in drug discovery? *Nat. Rev. Drug. Disc.* **8,** 865–878 (2009).

Akhondi, S. A. *et al.* Annotated chemical patent corpus: a gold standard for text mining. *PLOS One* **9,** e107477 (2014).

Anaya-Ruiz, M. & Perez-Santos, M. Innovation status of gene therapy for breast cancer. *Asian. Pac. J. Cancer. Prev.* **16,** 4133–4136 (2015).

Antunes, A., Fierro, I., Guerrante, R., Mendes, F. & Alencar, M. Trends in nanopharmaceutical patents. *Int. J. Mol. Sci.* **14,** 7016–7031 (2013).

Arshad, Z. *et al.* Open access could transform drug discovery: a case study of JQ1. *Expert. Opin. Drug. Discov.* **11,** 321–332 (2016).

Azoulay, P., Michigan, R. & Sampat, B. N. The anatomy of medical school patenting. *N. Engl. J. Med.* **357,** 2049–2056 (2007).

Bartolomé-Nebreda, J.-M., Conde-Ceide, S. & García, M. Phosphodiesterase 10A inhibitors: analysis of US/EP patents granted since 2012. *Pharm. Pat. Anal.* **4,** 161–186 (2015).

Bawa, R. Patents and nanomedicine. *Nanomedicine* **2,** 351–374 (2007).

Bergman, K. & Graff, G. D. The global stem cell patent landscape: implications for efficient technology transfer and commercial development. *Nat. Biotechnol.* **25,** 419–424 (2007)

Bijle, M. N. & Patil, S. Filed and granted Indian Patents in dentistry from 2005-2009: A critical analysis and review. *Indian J. Dent. Res.* **24,** 646 (2013).

Bonser, R. H. C. Patented biologically-inspired technological innovations: a twenty year view. *J Bionic. Eng.* **3,** 39–41 (2006).

Bowman, P. A. & Greenleaf, D. Non-CFC metered dose inhalers: the patent landscape. *Int. J Pharm,* **186,** 91–94 (1999).

Campos Jiménez, E. & Campos Ferrer, A. Analysis of immunological patents filed under the Patent Cooperation Treaty (2004–2011). *Inmunología* **33,** 21–26 (2014).

Canongia, C., Antunes, A. & Freitas Pereira, M. de N. Technological foresight—the use of biotechnology in the development of new drugs against breast cancer. *Technovation* **24,** 299–309 (2004).

Chandrasekharan, S. *et al.* Intellectual property rights and challenges for development of affordable human papillomavirus, rotavirus and pneumococcal vaccines: Patent landscaping and perspectives of developing country vaccine manufacturers. *Vaccine* **33,** 6366–6370 (2015).

Ciccone, M. *et al.* The role of omega-3 polyunsaturated fatty acids supplementation in childhood: a review. *Recent. Pat. Cardiovasc. Drug. Discov.* **8,** 42–55 (2013).

Clark, K. *et al.* Patent data mining: A tool for accelerating HIV vaccine innovation. *Vaccine* **29,** 4086–4093 (2011).

Cucoranu, I., Pantanowitz, L., Parwani, A., Vepa, S. & Weinstein, R. Digital pathology: a systematic evaluation of the patent landscape. *J Pathol. Inform.* **5,** 16 (2014).

da Silva Madeira, L., Borschiver, S. & Pereira Jr, N. Prospects and trends in the Brazilian market for biologically sourced products. *J. Technology Management & Innovation* **7,** 44–56 (2012).

Dalton, D. M., Burke, T. P., Kelly, E. G. & Curtin, P. D. Quantitative analysis of technological innovation in knee arthroplasty. *J. Arthroplasty* **31,** 1366–1372 (2016).

Dara, A. & Sangamwar, A. T. Clearing the fog of anticancer patents from 1993–2013: through an in-Depth Technology landscape & target analysis from pioneer research institutes and universities worldwide. *PLOS One* **9,** e103847 (2014).

Deiss, R. Intellectual property organizations and pharmaceutical patents in Africa. *Soc. Sci. Med.* **64,** 287–291 (2007).

Demunshi, Y. & Chugh, A. Patenting trends in marine bioprospecting based pharmaceutical sector. *J. Intellect. Prop. Rights* **14,** 122–130 (2009).

Deorsola, A. C., Mothé, C. G., de Oliviera, L. G. & Deorsola, A. B. Technological monitoring of cyclodextrin – World panorama. *World Pat. Inf.* **39,** 41–49 (2014).

Dou, H. & Bai, Y. A rapid analysis of Avian Influenza patents in the Espacenet® database – R&D strategies and country comparisons. *World Pat. Inf.* **29,** 26–32 (2007).

Emmerich, C. Comparing first level patent data with value-added patent information: A case study in the pharmaceutical field. *World Pat. Inf.* **31,** 117–122 (2009).

Erdin, N., Robin, F., Heinemann, L., Brandt, D. & Hovorka, R. Further development of artificial pancreas: blocked by patents? *J. Diabetes Sci. Technol.* **2,** 971–976 (2008).

Esmond, R. W. & Chung, A. K.-H. The patent landscape of siRNA nanoparticle delivery. *Nanotechnology Law & Business* **11,** 14 (2014).

Fechete, R. *et al.* Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. *Proteomics Clin. App.* **5,** 354–366 (2011).

Fiala, J. L. A. & Lowery, D. Migraine therapies targeting the CGRP pathway: intellectual property landscape: Patent watch. *Nat. Rev. Drug Disc.* **15,** 8–9 (2016).

Garrison, H. H., Herman, S. S. & Lipton, J. A. International distribution of dental materials publications and patents. *Dent. Mater.* **8,** 42–48 (1992).

Gaspar Amaral, L. F. & Fierro, I. M. Profile of medicinal plants utilization through patent documents: The andiroba example. *Rev. Bras. Farmacogn.* **23,** 716–722 (2013).

Gavaraskar, K., Dhulap, S. & Hirwani, R. R. Therapeutic and cosmetic applications of Evodiamine and its derivatives—A patent review. *Fitoterapia* **106,** 22–35 (2015).

Georgieva, B. P. & Love, J. M. Human induced pluripotent stem cells: a review of the US patent landscape. *Regen. Med.* **5,** 581–591 (2010).

Goetze, C. An empirical enquiry into co-patent networks and their stars: The case of cardiac pacemaker technology. *Technovation* **30,** 436–446 (2010).

Gupta, R. & Manchikanti, P. Analysis of patenting trends of antifungal drugs in the product patent regime in India. *World Pat. Inf.* **32,** 135–140 (2010).

Gupta, R. K. & Subbaram, N. R. Patenting activity in the field of biotechnology: Indian scenario. *World Pat. Inf.* **14,** 36–41 (1992).

Haanes, E. J. & Cànaves, J. M. Stealing fire: a retrospective survey of biotech patent claims in the wake of Mayo v. Prometheus. *Nat. Biotechnol.* **30,** 758–760 (2012).

Jana, T. *et al.* Antimalarial patent landscape: a qualitative and quantitative analysis. *Curr. Sci.* **103,** 14 (2012).

Jee, S. J. & Sohn, S. Y. Patent network based conjoint analysis for wearable device. *Technol. Forecast. Soc. Change* **101,** 338–346 (2015).

Jensen, K. & Murray, F. Intellectual property landscape of the human genome. *Science* **310,** 239–240 (2005).

Jun, S., Park, S. S. & Jang, D. S. Patent management for technology forecasting: A case study of the bio-industry. *J. Intellect. Prop. Rights* **17,** 539–546 (2012).

Kapczynski, A., Park, C. & Sampat, B. Polymorphs and prodrugs and salts (oh my!): an empirical analysis of "secondary" pharmaceutical patents. *PLOS One* **7,** e49470 (2012).

Kong, X., Hu, Y., Cai, Z., Yang, F. & Zhang, Q. Dendritic-cell-based technology landscape: Insights from patents and citation networks. *Hum. Vaccin. Immunother.* **11,** 682–688 (2015).

Konski, A. F. & Spielthenner, D. J. F. Stem cell patents: a landscape analysis. *Nat. Biotechnol.* **27,** 722–726 (2009).

Korting, H. C., Blecher, P., Schäfer-Korting, M. & Wendel, A. Topical liposome drugs to come: What the patent literature tells us. *J. Am. Acad. Dermato.* **25,** 1068–1071 (1991).

Lee, B. W. *et al.* Functional annotation and analysis of Korean patented biological sequences using bioinformatics. *Mol. Cells* **21,** 269-275 (2006).

Lee, E. C. Y. & Carpino, P. A. Melanocortin-4 receptor modulators for the treatment of obesity: a patent analysis (2008-2014). *Pharm. Pat. Anal.* **4,** 95–107 (2015).

Lim Chin, W. W., Parmentier, J., Widzinski, M., Tan, E. H. & Gokhale, R. A brief literature and patent review of nanosuspensions to a final drug product. *J. Pharm. Sci.* **103,** 2980–2999 (2014).

Lundin, P. Is silence still golden? Mapping the RNAi patent landscape. *Nat. Biotechnol.* **29,** 493–497 (2011).

Swamy, H. M. M., *et al.* Analysis of opportunities and challenges in patenting of bacillus thuringiensis insecticidal crystal protein genes. *Recent Pat. DNA Gene. Seq.* **6,** 64–71 (2012).

Ma, Z. *et al.* An assessment of traditional Uighur medicine in current Xinjiang region (China). *Afr. J. Tradit. Complement. Altern. Med.* **11,** 301 (2014).

Madeira, L. S., Borschiver, S. & Pereira, N. On the assignment of biopharmaceutical patents. *Technol. Forecast. Soc. Change* **80,** 932–943 (2013).

Mucke, H. A. M. Relating patenting and peer-review publications: an extended perspective on the vascular health and risk management literature. *Vasc. Health. Risk Manag.* **7,** 265–272 (2011).

Nascimento, M. L. F. Brief history of X-ray tube patents. *World Pat. Inf.* **37,** 48–53 (2014).

Padmanabhan, S., Amin, T., Sampat, B., Cook-Deegan, R. & Chandrasekharan, S. Intellectual property, technology transfer and manufacture of low-cost HPV vaccines in India. *Nat. Biotechnol.* **28,** 671–678 (2010).

Palazzoli, F., Bire, S., Bigot, Y. & Bonnin-Rouleux, F. Landscape of chromatin control element patents: positioning effects in pharmaceutical bioproduction. *Nat Biotechnol.* **29,** 593–597 (2011).

Palazzoli, F., Testu, F.-X., Merly, F. & Bigot, Y. Transposon tools: worldwide landscape of intellectual property and technological developments. *Genetica* **138,** 285–299 (2010).

Panja, S. *et al.* Global research on medical cotton - evidence from patent landscape study. *J. Intellect. Prop. Rights* **20,** 39–50 (2015).

Paradise, J., Andrews, L. & Holbrook, T. Intellectual property. Patents on human genes: an analysis of scope and claims. *Science* **307,** 1566–1567 (2005).

Petering, J., McManamny, P. & Honeyman, J. Antibody therapeutics – the evolving patent landscape. *New Biotechnol.* **28,** 538–544 (2011).

Piacentini, E., Drioli, E. & Giorno, L. Membrane emulsification technology: Twenty-five years of inventions and research through patent survey. *J. Memb. Sci.* **468,** 410–422 (2014).

Pierce, B. L., Carlson, C. S., Kuszler, P. C., Stanford, J. L. & Austin, M. A. The impact of patents on the development of genome-based clinical diagnostics: an analysis of case studies. *Genet. Med.* **11,** 202–209 (2009).

Roberts, M. *et al.* The global intellectual property landscape of induced pluripotent stem cell technologies. *Nat. Biotechnol.* **32,** 742–748 (2014).

Rodriguez-Monguio, R. & Seoane-Vazquez, E. Patent life of antiretroviral drugs approved in the US from 1987 to 2007. *AIDS Care* **21,** 760–768 (2009).

Rogueda, P., Lallement, A., Traini, D., Iliev, I. & Young, P. M. Twenty years of HFA pMDI patents: facts and perspectives: Twenty years of HFA pMDI Patents. *J. Pharm. Pharmacol.* **64,** 1209–1216 (2012).

Sahoo, N., Manchikanti, P. & Dey, S. H. Herbal drug patenting in India: IP potential. *J. Ethnopharmacol.* **137,** 289–297 (2011).

Sastry, K. R., Rashmi, H. B. & Badri, J. Research and development perspectives of transgenic cotton: evidence from patent landscape studies. *J. Intellect. Prop. Rights* **16,** 139–153 (2011).

Shahiwala, A., Vyas, T. & Amiji, M. Nanocarriers for systemic and mucosal vaccine delivery. *Recent Pat. Drug. Deliv. Formul.* **1,** 1–9 (2007).

Sternitzke, C. An exploratory analysis of patent fencing in pharmaceuticals: The case of PDE5 inhibitors. *Res. Policy* **42,** 542–551 (2013).

Sudhakar Johnson, T. *et al.* Genetic improvement of biofuel plants: recent progress and patents. *Recent Pat. DNA Gene. Seq.* **7,** 2–12 (2013).

Tang, X. & Du, J. Natural products against cancer: A comprehensive bibliometric study of the research projects, publications, patents and drugs. *J. Can. Res. Ther.* **10,** 27 (2014).

Telang, M. A., Bhutkar, S. P. & Hirwani, R. R. Analysis of patents on preeclampsia detection and diagnosis: A perspective. *Placenta* **34,** 2–8 (2013).

Uchôa, N. N., Ferreira, R. de P., Sachetto-Martins, G. & Müller, A. C. Ten years of the genomic era in Brazil: Impacts on technological development assessed by scientific production and patent analysis. *World Pat. Inf.* **33,** 150–156 (2011).

van Rooij, E., Purcell, A. L. & Levin, A. A. Developing microRNA therapeutics. *Circ. Res.* **110,** 496–507 (2012).

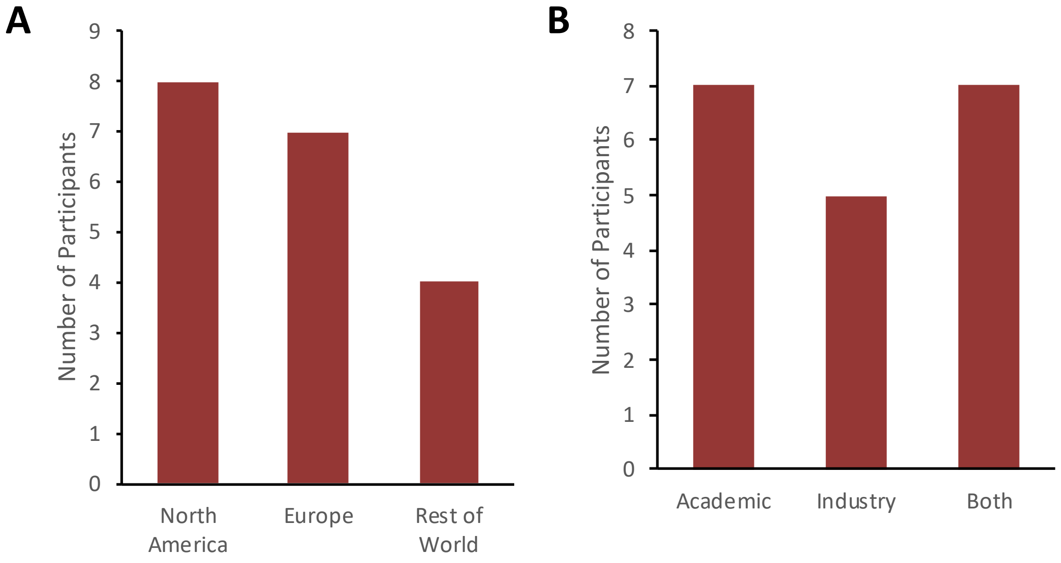
Verbeure, B., Matthijs, G. & Van Overwalle, G. Analysing DNA patents in relation with diagnostic genetic testing. *Eur. J. Hum. Genet.* **14,** 26–33 (2006).

Wang, X. Agricultural biotechnology worldwide patent analysis and mapping. *Afr. J. Biotechnol.* **10,** 1936–1944 (2011).

Wolfinbarger, L. Tissue engineering of bone and cartilage: a view through the patent literature. *Asian Biomed.* **5,** 1–12 (2011).

Xia, Y. *et al.* Literature and patent analysis of the cloning and identification of human functional genes in China. *Sci. China Life Sci.* **55,** 268–282 (2012).

8.2 Chapter 3



Supplementary Figure 8.1: Delphi study participant characteristics

A) Location and B) occupation of participants completing at least one round of the study (n = 19)

Supplementary Table 8.3: Results from round one of the modified Delphi study (n = 19)

Question Number	Question Item/ *Section*	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Range	Mode	SD
	Title									
1.	Authors should include that the article, or a component of the article, is a patent landscape in the title	3	8	9	8.4	10	10	5	10	1.4
	Summary/Abstract									
2.	Background to the study	5	8	9	8.5	10	10	5	10	1.8
3.	Rationale behind the study	4	8	10	8.8	10	10	6	10	1.9
4.	Aims of the study	8	10	10	9.8	10	10	2	10	0.5
5.	Sources of data for the patents included in the review	1	2	7	5.9	10	10	9	10	3.8
6.	Dates data sources were searched	1	2	4	4.9	8	10	9	2	3.4
7.	Patent offices searched	1	2	3	4.8	9	10	9	2	3.5
8.	Component(s) of patent documents searched	1	2	3	4.7	8	10	9	2	3.3
9.	Methods of analysis of data	1	3	6	6.2	9	10	9	10	3.0
10.	Results	9	10	10	9.9	10	10	1	10	0.2
11.	Main finding(s) in context of aim(s)	10	10	10	10.0	10	10	0	10	0.0
	Introduction									
12.	Aims of the study	6	10	10	9.6	10	10	4	10	1.0
13.	Rational behind the study	7	9	10	9.3	10	10	3	10	1.0

Question Number	**Question Item/ *Section***	**Min**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**	**Range**	**Mode**	**SD**
14.	Background to the study	10	10	10	10.0	10	10	0	10	0.0
15.	What the patent landscape adds to the literature	2	7	8	7.9	10	10	8	10	2.0
16.	Potential impact of the study	2	7	8	7.3	9	10	8	7	2.2
17.	For whom the review may be relevant	1	5	6	6.3	8	10	9	6	2.5
	Methods									
18.	State the subject matter of the patents aimed to be collected, e.g. “our search will collect patents regarding use of stem cells in treatment of melanoma”	2	9	10	8.9	10	10	8	10	2.0
19.	List the criteria used to select patents with reasoning	7	9	10	9.4	10	10	3	10	0.8
20.	Describe how patent documents were collected (e.g. through search of databases, or application of software).	1	7	9	8.1	10	10	9	10	2.7
21.	State the database(s) searched	9	9	10	9.7	10	10	1	10	0.5
22.	State patent office(s) searched	7	9	10	9.5	10	10	3	10	1.0
23.	State dates data sources were searched	5	10	10	9.3	10	10	5	10	1.6
24.	State component(s) of patent documents searched	5	9	10	9.2	10	10	5	10	1.4
25.	Search terms used for one database	3	8	10	8.4	10	10	7	10	2.3
26.	Search terms used for all databases searched	3	9	10	9.2	10	10	7	10	1.7
27.	Methods regarding how included patents were sorted for relevance; “Relevant patents were included through manual review of the title and abstract of each patent document”	2	7	9	8.5	10	10	8	10	2.0

Question Number	Question Item/ *Section*	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Range	Mode	SD
28.	State how the exclusion/ inclusion criteria were applied to each patent document (item removed due to similarity with 27)	1	3	7	6.2	9	10	9	3	2.9
29.	State that patent selection should be blindly reproduced	1	3	3	5.2	9	10	9	3	3.3
30.	State software used to extract data, if used	3	8	10	8.7	10	10	7	10	1.9
31.	State data collection was shown to be reproducible	1	3	5	5.6	8	10	9	3	3.1
32.	Describe all information sought from patent documents e.g. “Information regarding inventors and assignees was collected from each patent document”	3	8	9	8.3	10	10	7	9	2.0
33.	State what was done when this information was not available in a patent document	1	6	8	7.2	9	10	9	9	2.8
	Results									
34.	State number of patents assessed for eligibility	3	9	10	9.1	10	10	7	10	2.0
35.	State number of patents included in the study	5	10	10	9.5	10	10	5	10	1.2
36.	Provide reasons for exclusion at each stage	8	9	10	9.4	10	10	2	10	0.8
37.	Provide a summary of all patents included in the study	3	9	9	8.6	10	10	7	10	1.9
38.	Provide a list of patent publication numbers for patents included in the study.	1	3	5	5.8	8	10	9	5	3.0
39.	Present results of each analysis carried out with explanation of results.	5	9	10	9.1	10	10	5	10	1.6
40.	Provide an explanation of how each analysis contributes to the aims of the paper	2	8	9	8.0	10	10	8	10	2.3

Question Number	**Question Item/ *Section***	**Min**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**	**Range**	**Mode**	**SD**
41.	A list of patent publication numbers should be provided, for included patents, for each analysis carried out	1	2	5	5.6	9	10	9	5	3.2
	Discussion									
42.	Provide a summary of the studies findings	5	10	10	9.5	10	10	5	10	1.3
43.	Link the findings of the study to its aims	5	9	10	9.3	10	10	5	10	1.4
44.	Link the findings of the study to other work in the area	3	8	9	8.5	10	10	7	10	1.9
45.	Authors should discuss the potential impact of their work	3	8	8	8.4	10	10	7	10	2.0
46.	Authors should discuss to whom their work may be relevant	1	5	7	6.8	9	10	9	9	2.5
47.	Describe the limitations of the review in context of the reliability of the conclusions	3	9	10	9.1	10	10	7	10	1.7
48.	Discuss the limitations that are related to the methodology of the study	7	9	10	9.4	10	10	3	10	0.9
49.	Discuss the limitations that are related to the software used in the study	2	5	9	7.8	10	10	8	10	3.0
50.	Authors should discuss how they attempted to reduce sources of error	4	7	9	8.2	10	10	6	10	2.1
51.	Conclusion should provide a general interpretation of the results in the context of other evidence	3	8	10	9.1	10	10	7	10	1.7
52.	Conclusion should state how the work has implications for future research	5	8	9	8.8	10	10	5	10	1.5
53.	Conclusion should state how the study builds on previous	2	4	7	6.7	10	10	8	10	2.9

Question Number	Question Item/ *Section*	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Range	Mode	SD
	work									
	Other									
54.	Conflicts of Interest	6	10	10	9.6	10	10	4	10	1.0
55.	Describe sources of funding for the article	6	10	10	9.6	10	10	4	10	1.0
56.	Other support (e.g. supply of data)	6	10	10	9.4	10	10	4	10	1.2
57.	Describe the role of funders for the article	5	8	10	8.8	10	10	5	10	1.7

Supplementary Table 8.4: Results from round two of the modified Delphi study (n = 18)

Question Number	Question Item/ Section	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Range	Mode	SD
	Definition									
1	“NEW ITEM* Do you agree with this definition?	6	8	10	9.1	10	10	4	10	1.3
	Title									
2	*NEW ITEM* Authors should state the subject matter (e.g. stem cells for cancer, gene editing technologies) of the investigation in the title.	8	10	10	9.6	10	10	2	10	0.7
3	*NEW ITEM* Authors should state the main conclusion(s) in the title.	1	1	1	2.6	2	10	9	1	3.2
	Summary/Abstract									
4	Sources of data for the patents included in the review	2	3	3	3.7	3	10	8	3	2.2

Question Number	Question Item/ Section	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Range	Mode	SD
5	Dates data sources were searched	1	1	1	2.5	2.75	10	9	1	3.0
6	Patent offices searched	1	1	1	2.2	2	10	9	1	2.5
7	Component(s) of patent documents searched	1	1	1	1.9	1.75	10	9	1	2.3
8	Methods of analysis of data	2	3	3	3.8	3	10	8	3	2.2
9	*NEW ITEM* Funding information.	1	1	1	2.5	1	10	9	1	3.2
	Introduction									
10	What the patent landscape adds to the literature – Kappa is 0.67*	1	5.5	10	7.5	10	10	9	10	3.1
11	For whom the review may be relevant – Kappa is 0.69*	5	7.5	10	8.7	10	10	5	10	1.8
	Methods									
12	Describe how patent documents were collected (e.g. through search of databases, or application of software).	2	9.5	10	8.6	10	10	8	10	2.4
13	*NEW ITEM* Authors should specify the definition and source of patent family designations, e.g., Derwent or INPADOC, if any analysis incorporated patent families.	6	10	10	9.4	10	10	4	10	1.1
14	Methods regarding how included patents were sorted for relevance; “Relevant patents were included through manual review of the title and abstract of each patent document”	9	10	10	9.8	10	10	1	10	0.3
15	State that patent selection should be blindly reproduced	1	5	5	5.1	5	10	9	5	2.5
16	State data collection was shown to be reproducible	1	5	5	5.3	6	10	9	5	2.5
17	State what was done when this information was not available in a patent document	8	9	10	9.5	10	10	2	10	0.9

Question Number	**Question Item/ Section**	**Min**	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Max**	**Range**	**Mode**	**SD**
	Results									
18	Provide a list of patent publication numbers for patents included in the study.	2	8.5	10	8.5	10	10	8	10	2.3
19	A list of patent publication numbers should be provided, for included patents, for each analysis carried out.	1	1	1	2.4	2	10	9	1	2.9
	Discussion									
20	Authors should discuss to whom their work may be relevant	4	8.5	10	8.9	10	10	6	10	1.6
21	Discuss the limitations that are related to the software used in the study	3	9	10	8.9	10	10	7	10	1.7
22	Authors should discuss how they attempted to reduce sources of error	3	9	10	8.9	10	10	7	10	1.8
23	Conclusion should state how the study builds on previous work	2	9	10	8.3	10	10	8	10	2.7

8.3 Chapter 4

Supplementary Table 8.5: Search terms

Database Searched (Date Conducted)	**Search Terms**
Scopus (27/03/2018)	(TITLE-ABS-KEY (algorithm* OR predict* OR discriminat* OR differentiat* OR computat*) AND TITLE-ABS-KEY (drug* OR compound* OR molecule* OR "new chemical entit*" OR medication* OR medicine*) AND TITLE-ABS-KEY ("market approv*" OR "regulatory approv*" OR "market authori*" OR "regulatory authori*" OR "market launch" OR (drug W/1 approv*)) AND TITLE-ABS-KEY (fail* OR abandon* OR cease OR problem* OR unsuccess* OR discontinu*))
PubMed (9/04/2018)	(((((algorithm* or predict* or discriminat* or differentiat* or computat*))) AND (((drug* OR compound* OR molecule* or medication* or medicine*)) OR new chemical entit*)) AND (((((((market approv*) OR regulatory approv*) OR market authori*) OR regulatory authori*) OR market launch)) OR drug approv*)) AND ((fail* OR abandon* OR cease or problem* OR unsuccess* OR discontinu*))
MEDLINE (9/04/2018)	See Supplementary Table 8.6
EMBASE 1974 to present (28/08/2018)	((algorithm* or predict* or discriminat* or differentiat* or computat*) and (drug* or compound* or molecule* or medication* or medicine* or new chemical entit*) and (market approv* or regulatory approv* or market authori* or regulatory authori* or market launch or drug approv*) and (fail* or abandon* or cease or problem* or unsuccess* or discontinu*))

Supplementary Table 8.6: Search terms for MEDLINE through Ovid

Number	Searches	Results
1	*ALGORITHMS/ or *Pattern Recognition, Automated/	77,801
2	predict*.mp.	1,176,985
3	discriminat*.mp.	198,858
4	differentiat*.mp.	663,693
5	exp algorithms/	271,329
6	computat*.mp.	153,345
7	1 or 2 or 3 or 4 or 5 or 6	2,252,254
8	drug*.mp.	2,411,386
9	"new chemical entit*".mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	1,443
10	compound*.mp.	979,070
11	molecule*.mp.	595,071
12	medication*.mp.	255,555
13	medicine*.mp.	710,590
14	8 or 9 or 10 or 11 or 12 or 13	4,327,859
15	"market approv*".mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	239
16	"regulatory approv*".mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	1,791
17	"market authori*".mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	202
18	"regulatory authori*".mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	2,705
19	*Drug Approval/	6,080
20†	15 or 16 or 17 or 18 or 19	10,674

Number	Searches	Results
21†	7 and 14 and 20	510
22	fail*.mp.	1,027,346
23	abandon*.mp.	16,239
24	cease.mp.	3,589
25	problem*.mp.	831,112
26	unsuccess*.mp.	28,889
27	discontinu*.mp.	96,798
28	22 or 23 or 24 or 25 or 26 or 27	1,921,081
29†	21 or 28	1,921,495
30†	21 and 28	96
31	(drug adj1 approv*).mp. [mp=title, abstract, original title, name of substance word, subject heading word, keyword heading word, protocol supplementary concept word, rare disease supplementary concept word, unique identifier, synonyms]	15,346
32	15 or 16 or 17 or 18 or 19 or 31	19,680
33	7 and 14 and 28 and 32	191

† Not included directly in the final search, number 33

Supplementary Table 8.7: Data extraction items

Item/*Section*
General Information
Date of data extraction
Record number (e.g. PMID)
Lead author
Article title
Citation
Type of publication
Source of funding
Date of publication
Overall description of the approach

Dataset Characteristics

Source of data

Date dataset generated

Drug population (e.g. small molecules for cancer)

Definition of outcome variable

List and number of candidate predictor variables

Definition and source of candidate predictors provided?

Timing of prediction (e.g. Phase I, Phase II)

Sample size, including sample sizes of individual groups

Number of outcomes in relation to number of candidate predictors (events per variable)

Indications included

Date ranges for included drugs

Geographical limitations

Other restrictions on dataset

Number of entries with missing data, by outcome if available

Methods for Model Development, Performance and Evaluation

Handling of missing data

Description of modelling method

Final modelling method

Algorithm tuning/ development

Methods for selecting predictors for inclusion in multivariable modelling

Methods for selecting predictors during multivariable modelling

Treatment of continuous variables

Details of any sub-group analysis

Details of any comparator approach

Calibration measures (e.g. calibration plot, calibration slope)

Discrimination measures (e.g. area under the receiver operating characteristic curve)

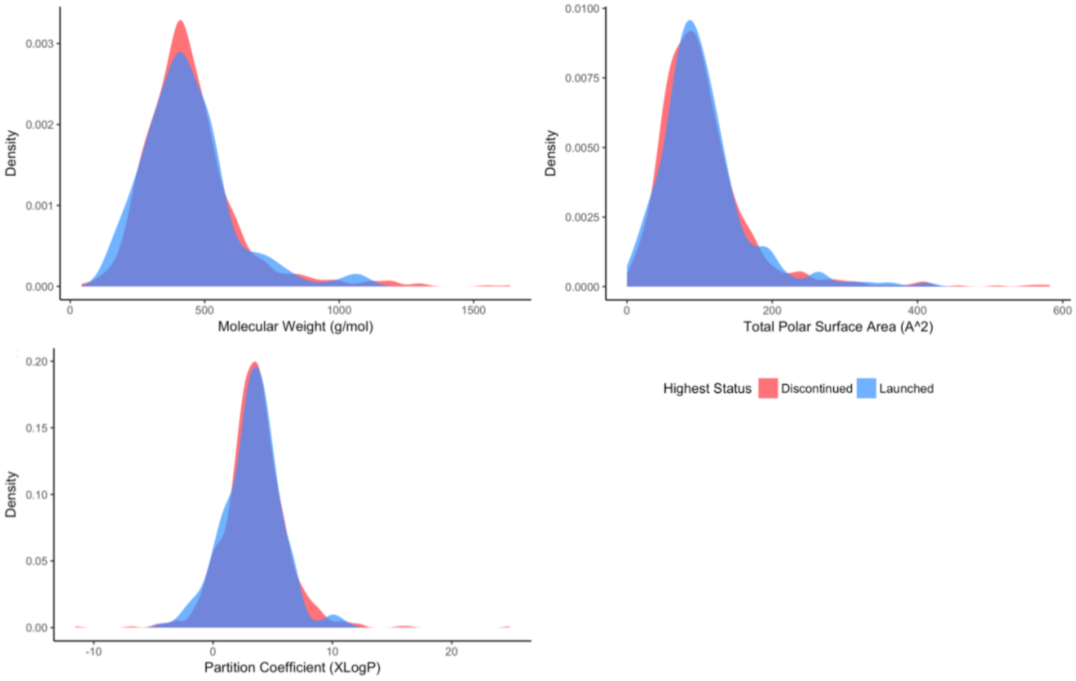
Classification measures (e.g. sensitivity, specificity)

Methods for testing model performance (e.g. cross-validation, bootstrapping, external validation)

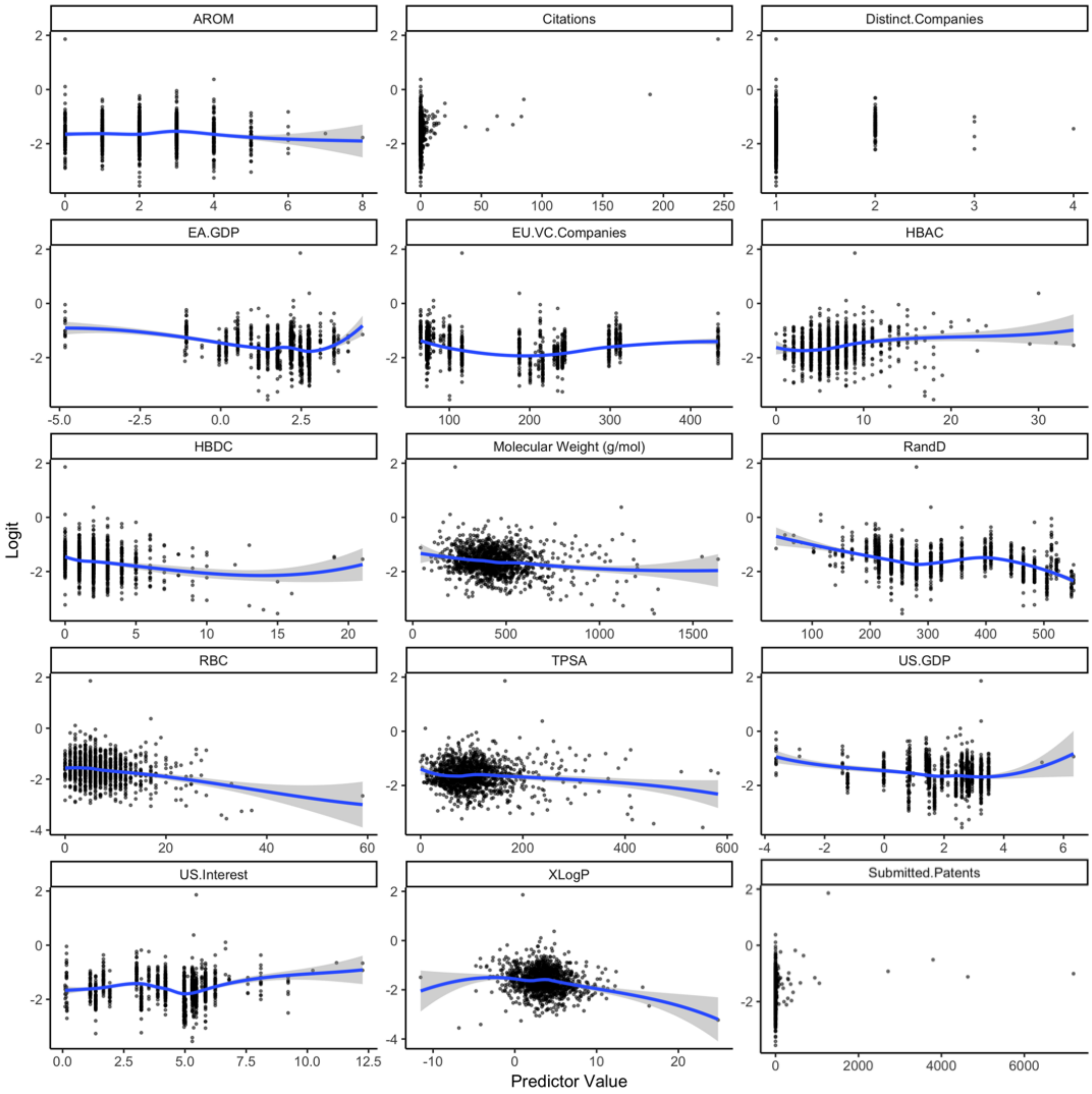
Results
Reported performance (discrimination, calibration, classification)
Details of final algorithm deployed
List of input variables required for algorithm
Details of relative importance or weight of input variables
Interpretation and Discussion
Limitations acknowledged in the manuscript
Other
Any other methodological issues identified

8.4 Chapter 5

8.4.1 Figures



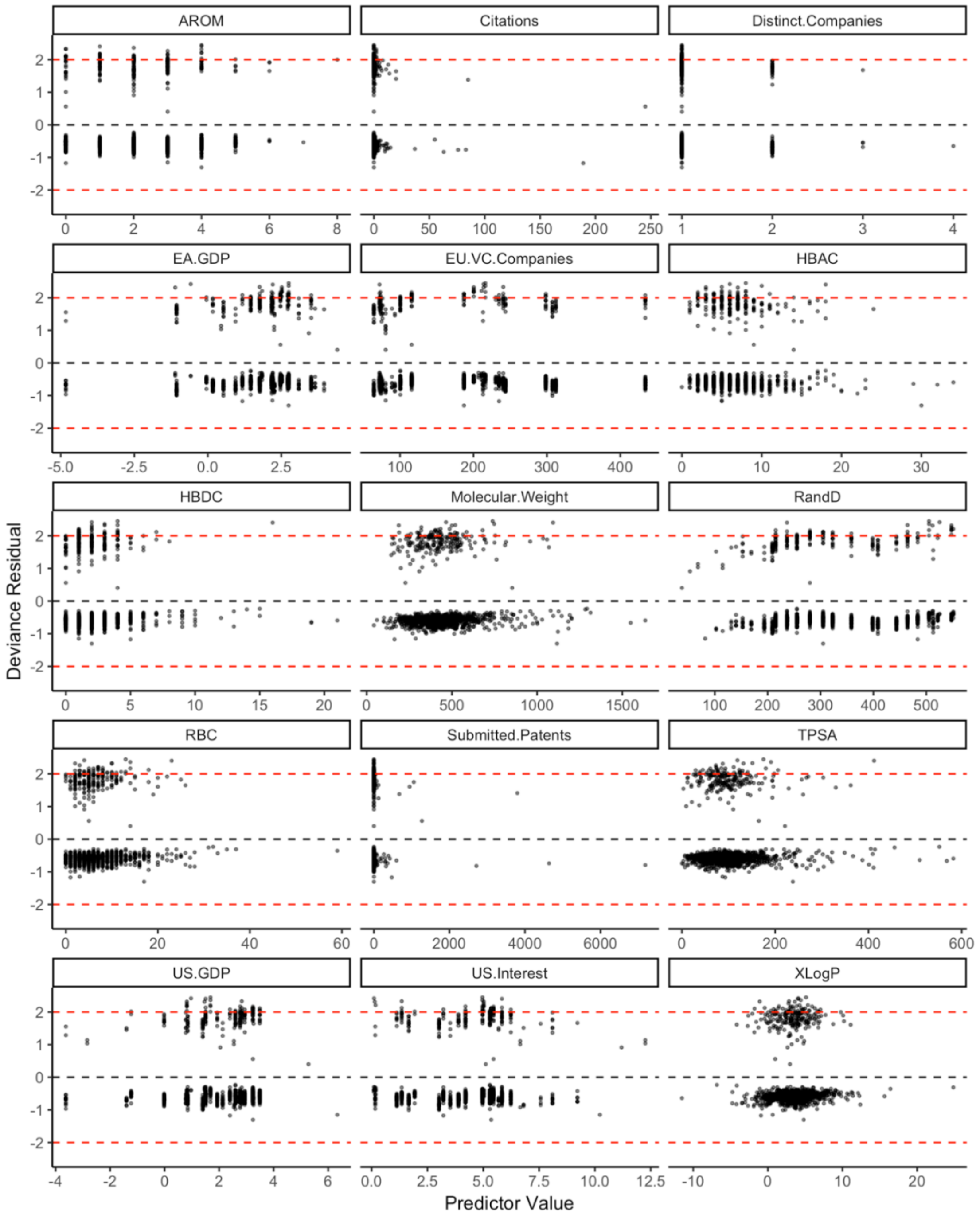
Supplementary Figure 8.2: Distributions of approved vs. discontinued drugs for example physicochemical parameters in the discovery dataset



Supplementary Figure 8.3: Checking the linearity in the logit assumption for a linear logistic regression model in the discovery dataset

Logistic regression assumes that predictor variables are linear in the logit. Numerical predictor variable values are plotted against the logit with a loess smoother (±95% CI). Where there were few distinct x-axis values and a large proportion of zeroes, a loess smoother was not fit. Abbreviations: AROM = number of aromatic rings, EA GDP = European Area gross domestic product change (annual %), EU VC Companies = number of companies receiving venture capital in EU (10s of companies) HBAC = hydrogen bond acceptor count HBDC =

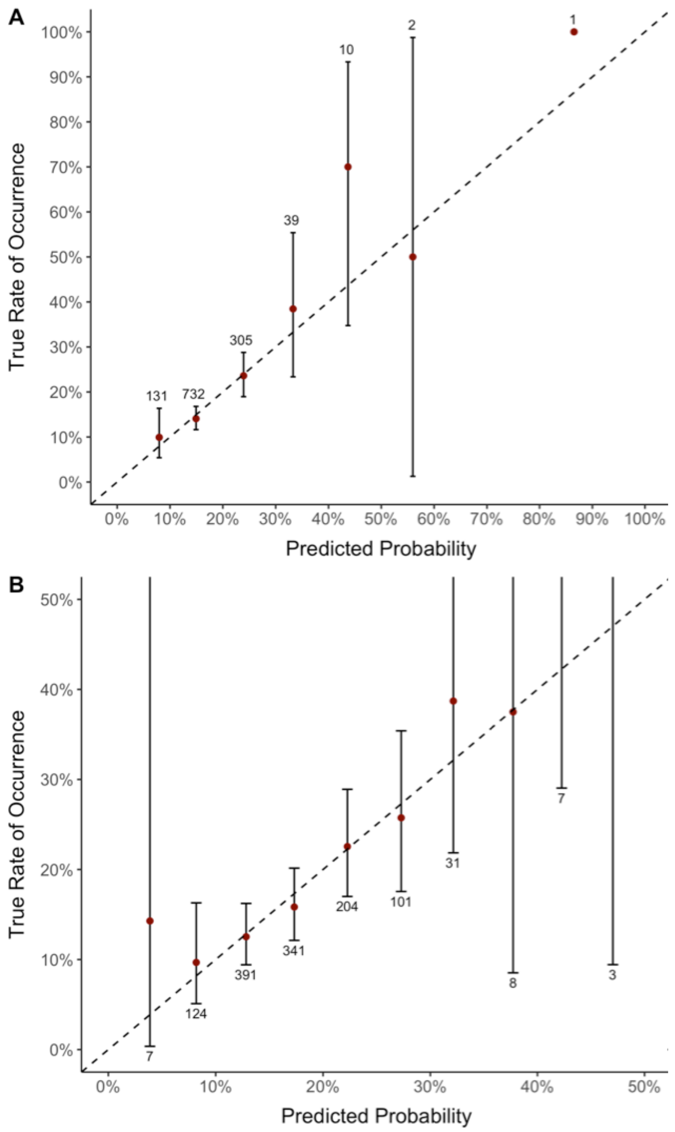
hydrogen bond donor count, RandD = pharma R&D spend, RBC = rotatable bond count, TPSA = topological polar surface area (Å^2), US GDP = United States gross domestic product change (annual %), XLogP = logarithm of the octanol/water partition coefficient (XLogP method). For full descriptions see Table 5.1 and methods.



Supplementary Figure 8.4: Deviance residuals for linear logistic regression on discovery dataset model show acceptable fit

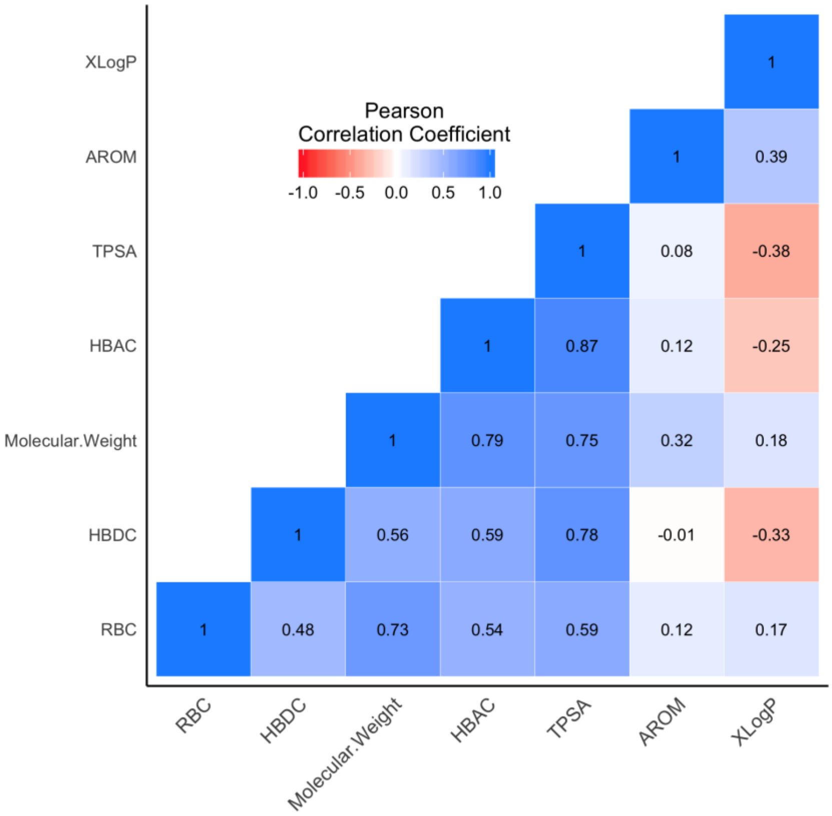
Most observations for numerical predictors yield a deviance residual within the range of ±2 (dashed red lines) indicating generally acceptable fit. Dashed black lines are zero (perfect fit).

Abbreviations: AROM = number of aromatic rings, EA GDP = European Area gross domestic product change (annual %), EU VC Companies = number of companies receiving venture capital in EU (10s of companies) HBAC = hydrogen bond acceptor count HBDC = hydrogen bond donor count, RandD = pharma R&D spend, RBC = rotatable bond count, TPSA = topological polar surface area (Å^2*), US GDP = United States gross domestic product change (annual %), XLogP = logarithm of the octanol/water partition coefficient (XLogP method). For full descriptions see Table 5.1 and methods.*



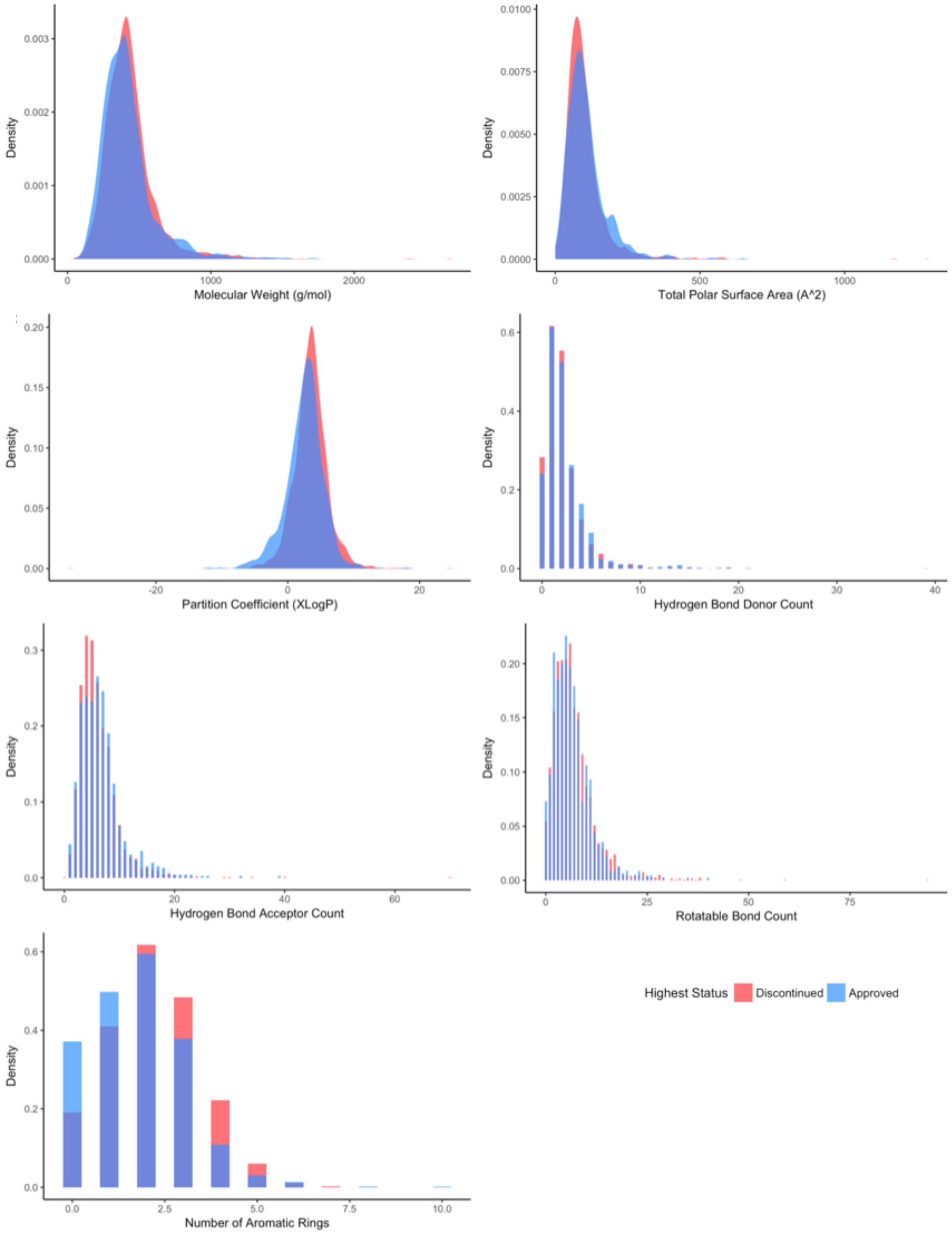
Supplementary Figure 8.5: Apparent calibration plots for discovery model

The discovery model was used to generate predicted probabilities for each drug the discovery dataset. Event-prediction pairs were then grouped into bins according to the predicted probabilities and plotted (error bars ± 95% CI). Number of observations per bin are shown above or below error bars. A) All predictions across the full range of predicted probabilities are shown (10 bins); B) Only predicted probabilities less that 50% are shown (20 bins across all data). Dashed line is perfect calibration.

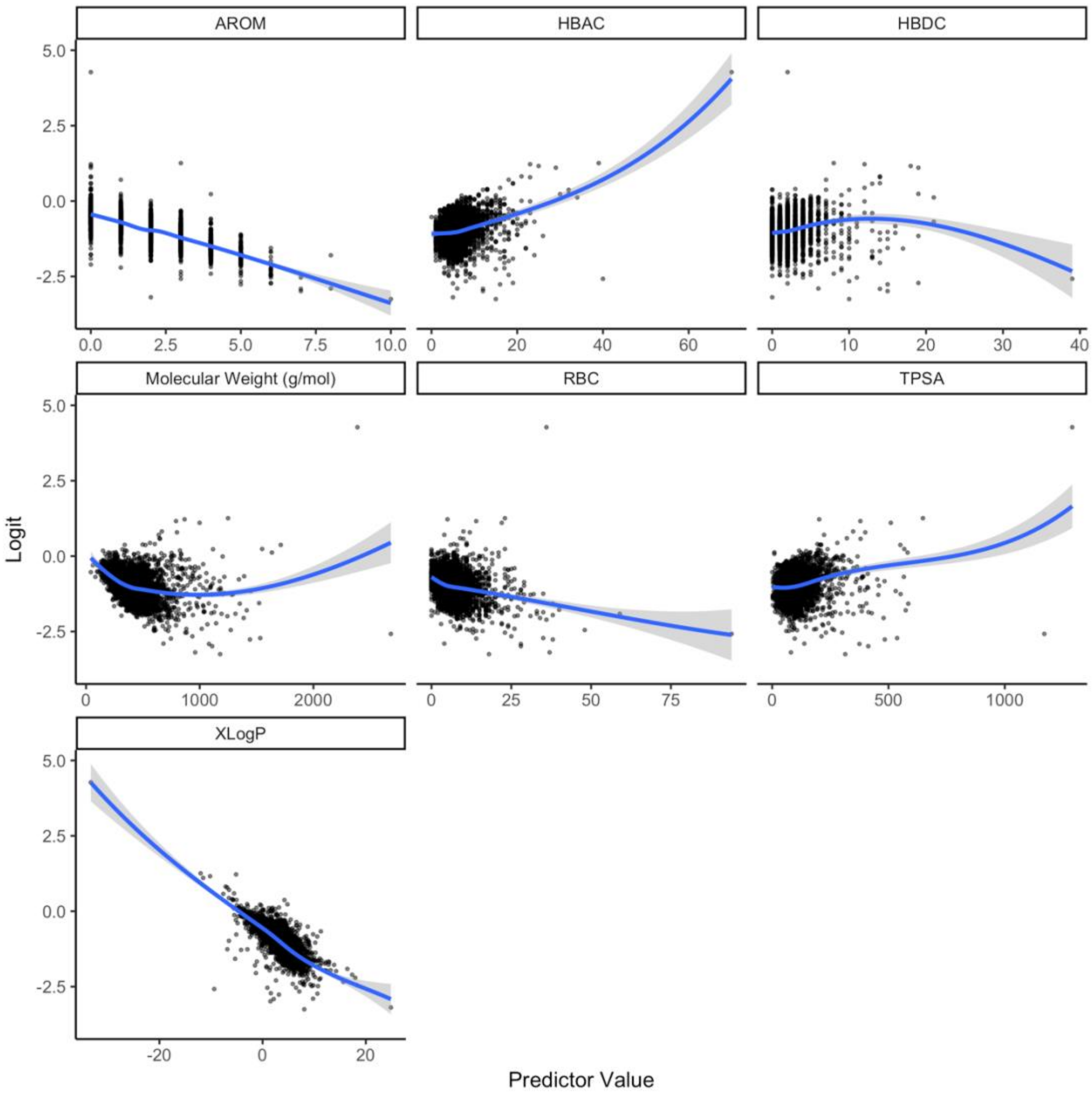


Supplementary Figure 8.6: Correlation matrix for predictor variables in physicochemical model

A correlation matrix was generated to allow assessment of collinearity, which was found to be acceptable. Abbreviations: AROM = number of aromatic rings, HBAC = hydrogen bond acceptor count HBDC = hydrogen bond donor count, RBC = rotatable bond count, TPSA = topological polar surface area, Å^2, XLogP = logarithm of the octanol/water partition coefficient (XLogP method).

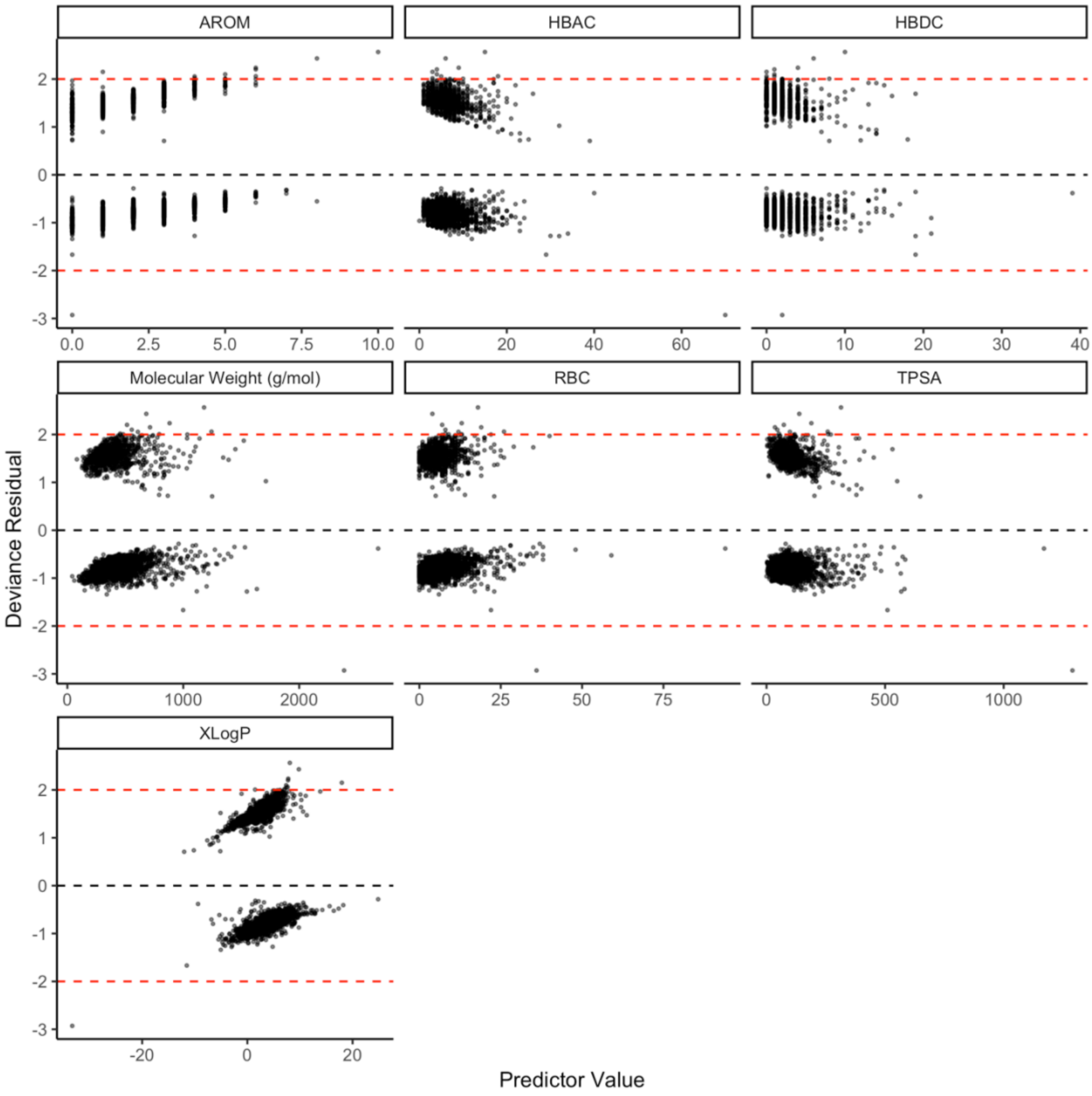


Supplementary Figure 8.7: Distribution of predictors in physicochemical dataset across approved and discontinued drugs



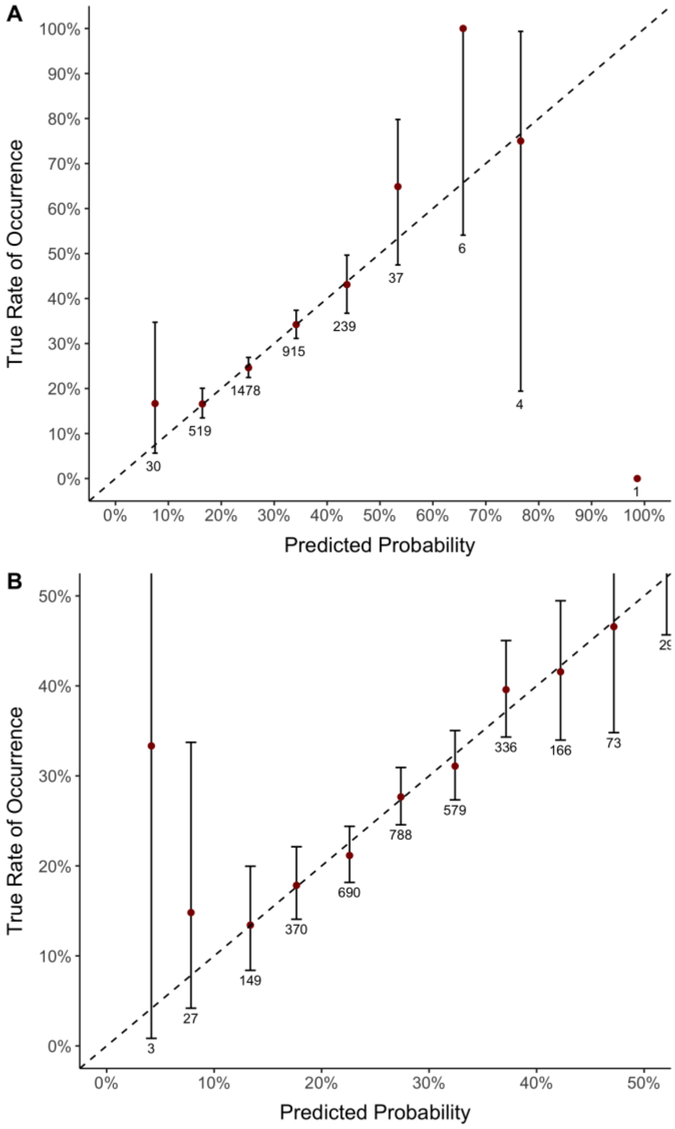
Supplementary Figure 8.8: Checking the linearity in the logit assumption for a linear logistic regression model in the physicochemical dataset

Logistic regression assumes that predictor variables are linear in the logit. Predictor values with a loess smoother (± 95% CI) are plotted against the logit and satisfy this assumption. Abbreviations: AROM = number of aromatic rings, HBAC = hydrogen bond acceptor count HBDC = hydrogen bond donor count, RBC = rotatable bond count, TPSA = topological polar surface area, Å^2, XLogP = logarithm of the octanol/water partition coefficient (XLogP method).



Supplementary Figure 8.9: Deviance residuals vs. predictor value for physicochemical model

Nearly all observations yield a deviance residual within the range of ±2 (dashed red lines) indicating good model fit. Dashed black lines are zero (perfect fit). Abbreviations: AROM = number of aromatic rings, HBAC = hydrogen bond acceptor count HBDC = hydrogen bond donor count, RBC = rotatable bond count, TPSA = topological polar surface area, Å^2, XLogP = logarithm of the octanol/water partition coefficient (XLogP method).



Supplementary Figure 8.10: Apparent calibration plots for physicochemical model

The physicochemical model was used to generate predicted probabilities for each drug in the physicochemical dataset. Event-prediction pairs were then grouped into bins according to the predicted probabilities and plotted (error bars ± 95% CI; not shown if only one observation). Number of observations per bin are shown above or below error bars. A) All predictions across the full range of predicted probabilities are shown (10 bins); B) Only predicted probabilities less that 50% are shown (20 bins across all data). Dashed line is perfect calibration.

8.4.2 Tables

Supplementary Table 8.8: Comparison of regression coefficients with and without influential observations in discovery dataset

Two influential observations were identified by Cook's distance calculations (observation 272 and 569). The impact of their removal on the regression coefficients is therefore presented.

Intercept and Predictor*	**All Observations**	**272 Removed**		**569 Removed**	
	Coefficient	Coefficient	% Change	Coefficient	% Change
(Intercept)	-0.077	-0.141	-84	-0.119	-54
Molecular weight	-0.002	-0.002	1	-0.002	4
HBDC	-0.070	-0.064	8	-0.069	1
HBAC	0.160	0.162	-1	0.159	1
RBC	-0.011	-0.011	4	-0.013	-15
TPSA	-0.002	-0.002	-17	-0.002	0
AROM	0.039	0.038	2	0.039	0
XLogP	0.014	0.013	4	0.014	-4
Academic commercial (commercial = 1)	-0.149	-0.151	-1	-0.150	-1
Company change (change = 1)	0.420	0.404	4	0.419	0
Distinct companies	-0.026	-0.013	50	-0.026	1
Citations	0.010	0.019	-90	0.009	9
Submitted patents	0.000	0.000	14	0.000	-176
RandD	-0.004	-0.004	3	-0.004	1
US interest rate	-0.042	-0.037	12	-0.040	7
EU VC companies	0.002	0.002	3	0.002	0
EA GDP	-0.106	-0.112	-6	-0.107	-1
US GDP	-0.066	-0.062	7	-0.065	2

**Abbreviations: AROM = number of aromatic rings, EA GDP = European Area gross domestic product change (annual %), EU VC Companies = number of companies receiving venture capital in EU (10s of companies) HBAC = hydrogen bond acceptor count HBDC = hydrogen bond donor count, RandD = pharma R&D spend, RBC = rotatable bond count, TPSA = topological polar surface area (Å^2), US GDP = United States gross domestic product change (annual %), XLogP = logarithm of the octanol/water partition coefficient (XLogP method). For full descriptions see Table 5.1 and methods.*

Supplementary Table 8.9: Comparison of regression coefficients with and without influential observation in the physicochemical dataset

One influential observation was identified by Cook's distance calculations (observation 666). The impact of its removal on the regression coefficients is therefore presented.

Intercept and Predictor*	Coefficient		Percentage Change
	All Observations	Influential Observation Removed	
(Intercept)	-0.403	-0.430	7
Molecular.Weight	-0.001	-0.001	-20
HBDC	-0.012	-0.042	258
HBAC	0.112	0.111	-1
RBC	-0.019	-0.020	2
TPSA	-0.002	-0.001	-40
AROM	-0.216	-0.219	1
XLogP	-0.056	-0.067	20

** Abbreviations: AROM = number of aromatic rings, HBAC = hydrogen bond acceptor count HBDC = hydrogen bond donor count, RBC = rotatable bond count, TPSA = topological polar surface area, Å^2, XLogP = logarithm of the octanol/water partition coefficient (XLogP method).*