Mapping the boundaries of medical knowledge using machine learning algorithms applied to the Medline database of peer reviewed medical literature to guide decision-makers during curriculum development and maintenance

Andrew Amos

B.Commerce – University of Adelaide, Australia

B.Arts (Honours) – University of Adelaide, Australia

MB.BS – University of Queensland, Australia

Submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

At the College of Medicine and Dentistry

James Cook University

February, 2026

Statement of Access

I, Andrew Amos, the author of this thesis, understand that this thesis will be made available for use to others. All users consulting this thesis will have to sign the following statement: In consulting this thesis I agree not to copy or closely paraphrase it in whole or part without the written consent of the author; and to make proper public written acknowledgement for any assistance which I have obtained from it.

Beyond this, I do not wish to place any restriction on access to this thesis.

……. 12.02.2026

(Author's signature) (Date)

Declarations

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text, and a list of references is given.

I declare that the research included in this thesis was ethically conducted, with approval received for each component of the research from the James Cook University Human Research Ethics Committee. The following ethics approval codes relate to the research relevant to this thesis: H9432.

……. 12.02.2026

(Author's signature) (Date)

Acknowledgements

I would like to express my gratitude to my primary advisor, Professor Bunmi Malau-Aduli, who gracefully guided me through the thickets and away from (sometimes out of) the cul-de-sacs that threaten progress in the journey of every thesis. Her patience, wisdom, and knowledge contributed throughout all phases of the journey.

I would also like to thank my secondary advisors, Professor Tarun Sen Gupta and Dr Joanne (Kyungmi) Lee. Tarun's great knowledge of and experience in curriculum development were particularly useful in grounding this very theoretically driven thesis in practical realities, while Joanne's programming expertise was particularly helpful in developing and explaining the technical features of the research.

Statement of the Contribution of Others

I would like to thank the following people for their support in the following roles leading to the completion of this thesis.

Supervision

Professor Bunmi Malau-Aduli

Professor Tarun Sen Gupta

Dr Joanne (Kyungmi) Lee

Editorial Support

Professor Bunmi Malau-Aduli

Professor Tarun Sen Gupta

Dr Joanne (Kyungmi) Lee

Research Assistance

Professor Bunmi Malau-Aduli

Professor Tarun Sen Gupta

Dr Joanne (Kyungmi) Lee

Thesis Abstract

Introduction: Competent medical practice depends on the acquisition and maintenance of a minimum set of knowledge, skills, and attributes to provide safe, ethical medical treatment to patients, and to uphold the other professional standards required in the various roles performed by doctors. Medical curricula are the frameworks that define what the minimum set of knowledge, skills, and attributes should be for generalists and specialists, and at different stages of training and practice.

Among the challenges to medical curriculum development is the enormous volume of medical research, as well as the accelerating rate of new research published each year. No individual human expert can read or understand even a tiny fraction of the new research, leading to practice silos with experts in one or two of the many potential clinical, administrative, and basic science domains relevant to medical practice.

The biases of medical experts are known to have contributed to significant harms to particular groups, including women, but the lack of empirical tools to understand the structure and organisation of the many domains of medical knowledge means that the content, structures, and priorities of medical curricula rely entirely upon expert consensus.

This research was designed to develop and begin to validate a comprehensive map of the knowledge contained within the Medline database of peer-reviewed medical literature, as a first step towards providing an empirical basis for medical curriculum development.

Methods: The thesis combined results from a systematic review of evidence for biases in empirical methods of selection into medical specialist training, computational modelling, and an online questionnaire of practicing psychiatrists.

A self-organising map called MedSOM was trained on the complete set of articles and medical subject headings indexed by the Medline database on 01/01/2021 and used to differentiate the

knowledge structures associated with medical and psychiatric practice. MedSOM was then used to analyse the clustering of psychiatric knowledge across ten editions of the core psychiatric textbook *Kaplan and Sadock's Comprehensive Textbook of Psychiatry* (*KSCTP*). A separate article described the technical innovations required to expand the SOM from a subset to the entire set of Medline articles.

Next, the bibliometric Increment statistic was applied to all Medline indexed articles relevant to psychiatric or psychological practice to identify topics of emerging interest. A co-word analysis of the top 20 topics was completed for each of 9 five-year periods (demi-decades) between 1972-1976 and 2012-2016, and the results analysed for coherence with events in psychiatric/psychological history over those periods.

To validate that emerging topics identified by the Increment statistic were meaningful to human experts, an online questionnaire asked a sample of 21 practicing psychiatrists to rank the novelty and practical relevance of 5 articles selected at random and 5 articles addressing emerging topics.

Results and discussion:

The systematic review found limited evidence that empirical methods of selecting junior doctors into specialist medical training programs were the cause of the under-representation of some groups in specialist training and practice. It found no evidence of techniques that successfully increased the selection of currently under-represented groups. Confidence in the results was low due to multiple methodological shortcomings, particularly the lack of reported power analyses, small sample sizes, lack of replication of findings, and the lack of a consensus in the literature regarding the results.

The technical innovations successfully allowed for the expansion of the SOM training corpus from a subset of 2 million Medline articles to the entire set of more than 30 million available at 01/01/21, and to consider all 29,917 medical subject headings instead of the 2,300 included in previous research. Ongoing technical innovations are likely to be necessary given the ever-expanding set of medical research articles, the addition of new medical subject headings, and increasing the utility of the SOM model by adding the capacity to model the effects of time.

MedSOM coherently clustered the knowledge represented by the reference lists of ten editions of *KSCTP*, and the evolution of the clusters over time was organised with reference to broad categories of psychiatric practice such as Adult Psychiatry, Child Psychiatry, and Administrative Psychiatry. The coherence of the organisation of the textbook, a standard source of psychiatric knowledge, with MedSOM provide support that MedSOM has extracted meaningful information about psychiatric knowledge from articles indexed in the Medline database.

Emerging topics among 18 million articles identified in 9 demi-decades between 1972-1976 and 2012-2016 were consistent with historical psychiatric/psychological milestones such as the publication of subsequent editions of the Diagnostic and Statistical Manual of Mental Disorders. Visualisations of the organisation of emerging topics in each demi-decade identified time-relevant themes, such as the recent focus on psychological and social factors implicated in suicide and suicide prevention.

Twenty-one psychiatrists ranked 5 articles selected by a machine learning algorithm as addressing emerging topics that were significantly more novel ($p<.03$) and more relevant ($p<.01$) than 5 articles selected at random. Novelty and relevance were not closely correlated. This was interpreted as preliminary support for the ability of the algorithm to identify features that have meaning to practicing psychiatrists, therefore supporting the validity of the algorithm's output.

Conclusions: This research has taken the first steps towards the development of an empirically derived map of medical knowledge capable of facilitating medical curriculum development. One of the anticipated advantages of an empirically derived map is to provide an unbiased standard against which to identify and address biases that arise from the reliance upon expert judgement in the construction of medical curricula.

The thesis illustrated that MedSOM could summarise the entire set of Medline indexed articles, analysing structural features such as the differentiation between medical and psychiatric knowledge, and analysing the structure of other forms of medical authority such as a core psychiatric textbook.

The thesis also provided preliminary evidence of the validity of the outputs of MedSOM, and the emerging topics algorithm, based on their coherence with milestones in psychiatric/psychological history; and the demonstrated capacity to identify articles that were significantly more novel and relevant to practice than articles selected at random.

Having established the equivalent of an outline of the primary towns and countries of a political map for the concepts and categories of the domain of medical knowledge, further steps towards a useful map will require close attention to a hybrid epistemology that considers the meaning of machine learning algorithms within the computational framework as well as the meaning inferred by human experts trying to make use of them.

List of Publications

Publication-Based Thesis Chapters

1. Amos A, Lee K, Sen Gupta T, Malau-Aduli BS. Systematic review of specialist selection methods with implications for diversity in the medical workforce. *BMC Medical Education*. 2021; 21:448. https://doi.org/10.1186/s12909-021-02685-w
2. Amos A, Lee K, Sen Gupta T, Malau-Aduli BS. Identifying emerging topics in the peer-reviewed literature to facilitate curriculum renewal and development. *Current Psychology*. 2023; 42:30813-30824. https://doi.org/10.1007/s12144-022-04090-y
3. Amos A, Lee K, Sen Gupta T, Malau-Aduli BS. Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps. *Studies in Health Technology and Informatics.* 2024; 310:795-799. https://doi.org/10.3233/SHTI231074
4. Amos A, Lee K, Sen Gupta T, Malau-Aduli BS. Validating the knowledge represented by a self-organizing map with an expert-derived knowledge structure. *BMC Medical Education.* 2024; 24(1): 416. https://bmcmededuc.biomedcentral.com/articles/10.1186/s12909-024-05352-y
5. Amos A, Lee K, Sen Gupta T, Malau-Aduli BS. Testing the validity of emerging topics in psychiatry identified by a machine learning algorithm with a sample of working psychiatrists. *Artificial Intelligence.* (Submitted)
6. Amos A, Lee K, Sen Gupta T, Malau-Aduli BS. Novel sparse matrix algorithm expands the feasible size of a self-organizing map of the knowledge indexed by a database of peer-reviewed medical literature. *Computing*. (Submitted)

Conference Presentations

1. Amos A. Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps. Oral presentation at the 19th World Congress on Medical and Health Informatics (MEDINFO 2023) at the International Convention Centre (ICC) in Sydney, Australia on 11th July, 2023.

Table of Contents

List of Tables

List of Figures

Chapter 5: Identifying emerging topics in the peer-reviewed literature to facilitate curriculum renewal and development

Chapter 6: Novel sparse matrix algorithm expands the feasible size of a self-organizing map of the knowledge indexed by a database of peer-reviewed medical literature

Chapter 7: Validating the knowledge represented by a self-organizing map with an expert-derived knowledge structure

List of Abbreviations

AI	Artificial intelligence
ARCP	Annual review of competence progression
BSOM	Batch self-organising map
CME	Continuing medical education
CPD	Continuing professional development
CPST	Clinical problem-solving test
DMG	Domestic medical graduate
DSM	Diagnostic and Statistical Manual of Mental Disorders
FSMB	Federation of State Medical Boards
GIS	Geographic information system
GPU	Graphics processing unit
IMG	International medical graduate
KSA	Knowledge, skills, and attributes
KSCTP	Kaplan and Sadock's Comprehensive Textbook of Psychiatry
LBD	Literature-based discovery
LibSVM	Library for support vector machines
LoR	Letter of recommendation
MB	Megabyte
MCQ	Multiple choice question
MedSOM	Medical self-organising map
MERSQI	Medical Education Research Study Quality Instrument
MeSH	Medical subject heading
ML	Machine learning
MMI	Multiple mini-interviews

MSRA	Multi-Specialty Recruitment Assessment
MSTS	Medical specialty training selection
NBME	National Board of Medical Examiners
NfL	Neurofilament light
NLM	National Library of Medicine
PLAB	Professional and linguistic assessment board
PMID	PubMed ID
RANZCP	Royal Australian and New Zealand College of Psychiatrists
ReDSOM	Recurrent density self-organising map
RNN	Recurrent neural network
SciSci	Science of Science
SJT	Situational judgement test
SOM	Self-organising map
SoS	Science of Science
STAR	Selection Tool for Applicants to Residency
URM	Under-represented minorities
USMLE	United States Medical Licensing Exam
WoS	Web of Science
WPA	World Psychiatric Association
xAI	Explainable artificial intelligence

Chapter 1: General Introduction

This thesis describes a program of research designed to take the first steps towards an empirically derived map of medical knowledge suitable to guide medical curriculum development. This chapter outlines the background knowledge necessary to understand the aims and hypotheses which generated the research, from the epistemology of medical curricula, the roles of theory, evidence, and expert knowledge in their creation, and the potential of machine learning (ML) algorithms to address problems with curriculum development such as expert bias.

After stating the aims and hypotheses of the research, the chapter describes the theoretical framework, methodologies, expected outcomes, and structure of the thesis.

1.1 - Epistemological underpinnings of medical curriculum development

Medical curricula define the set of knowledge, skills, and attributes (KSA) required for competent medical practice, the experiences by which these will be acquired, and how their acquisition and maintenance will be confirmed.[1] The content and structure of current medical curricula depend almost entirely upon expert judgement shaped by consensus methods such as the Delphi technique.[2] While curricula defined by medical experts have face validity, empirical evidence shows that expert consensus hides biases that can negatively affect large groups of patients.[3] This thesis is designed to explore the possibility that recently developed ML techniques which condense enormous datasets of medical knowledge and practice into empirical evidence will be able to inform curriculum development and help address expert limitations including bias.[4]

Medical education exists on a continuum from the more prescriptive programmes used with medical students studying for university degrees, through generalist/specialist training programmes for junior doctors developing expertise, to continuing professional development (CPD) frameworks designed to maintain and expand the capacities of practising consultants in specialised areas of practice.[5] Currently, there are no widely used empirical bases for establishing the boundaries of generalist or

specialist medical KSA, the currency and clinical importance of different elements of KSA, or the relative importance of these elements as features of a training curriculum.

In addition, there is almost no formal treatment of epistemology during medical education, training, and practice. The most common epistemological model used by practicing doctors is evidence-based medicine, but generally it is understood as a loose set of heuristics applied unsystematically to address specific clinical questions. It has been suggested that uncritical use of aggregated evidence in systematic reviews and clinical guidelines can lead to the same sort of treatment based on appeals to authority, without understanding, that evidence-based medicine was intended to replace.[6]

1.2 - The integration of empirical evidence, expert opinion, and pedagogical theory in medical curriculum development

Ideally, medical curricula would be the result of expert judgement informed by theory and empirical evidence.[2] Theories, such as models of the mechanisms of learning in adult professionals, can provide frameworks within which to understand students' and doctors' experiences of education, teaching, and practice. Empirical evidence provides data against which theories can be tested, outcomes can be projected, and expectations can be established and then monitored. Despite the potential advantages, theories of learning and empirical evidence about the KSA required for medical practice play a minor role in modern curriculum development and maintenance.[7,8]

In particular, despite the importance of a well-designed curriculum based on the best available evidence for establishing a feasible and equitable program balancing the needs of students, doctors, patients, and society,[1,9] there is little evidence regarding methods used to select the content included in, and explicitly or implicitly excluded from, medical education curricula. Most curriculum development initiatives described in peer reviewed publications use consensus-based methods such as Delphi for content selection. Such methods attempt to minimise bias by integrating a broad range of opinions rather than by drawing on objective evidence about what content is necessary and sufficient, but there is little evidence that they achieve this goal.[2]

One of the core functions of medical curricula is to specify the limits and level of mastery of KSA expected at each level of progression through medical training, including at entry to medical school. Kane's approach to the validity of assessments focuses on the nature of the decisions being made based on those assessments.[10,11] The content included in/excluded from medical education curricula influence decisions ranging from which candidates are selected into medical school, through graduation and certification as junior doctors, to entry into post-graduate training and progression to independent practice as a qualified consultant.

Outside training, content decisions will also have an impact on the KSA possessed, valued, and practiced by doctors, and thereby impact patient outcomes.[12] Biased selection of the content included in medical curricula may contribute to or prolong poor patient outcomes, for example, the gender biases in cardiac research which are argued to have caused suboptimal outcomes for women for decades.[13] As a result, adequate content validity, which is the extent to which the items on a test are fairly representative of the entire domain the test seeks to measure, is a crucial characteristic of medical curricula.

Recognition of the limitations of consensus-based curriculum development has led to calls for the development of empirical methods for determining the content of medical curricula,[14] but the literature does not report widespread use of such methods.[2] In their absence medical curriculum content selection decisions are entirely dependent on expert judgement. The sheer volume of medical knowledge and the increasing rate of its production has made it difficult to produce coherent empirically derived descriptions of the overall structure and relative importance of the domains of KSA from which experts select content for medical curricula.[15]

1.3 - Medical curricular biases and the under-representation of women and ethnic minorities across the career span

While the balance is changing, senior medical roles in developed Western countries have been dominated by males from the dominant ethnic group(s) of each country, including in the academic

roles that create and administer medical training programmes.[16] As a result, it is likely that the content of curricula for the education of medical students and specialist training of medical registrars has also mostly been selected by European males, although this has not been empirically confirmed.

At the same time, curricula for the education of medical students and specialist training of medical registrars in these countries have been shown to contain gender and racial biases that disadvantage women and minorities.[17,18] Although no causal link has been proven, it has been hypothesised that ethnocentric and male domination of senior academic roles has played a part in the under-representation of some groups at all levels of the medical career, from selection into medical school through to promotion to senior roles.[19]

Among the proposed causes of under-representation are the implicit biases of senior academic doctors who unfairly evaluate women and minorities seeking promotion who do not share their knowledge, experiences, and priorities.[20,21] These biases can be directly expressed in the content selected for inclusion in medical curricula. For example, Arsever and colleagues reviewed the clinical vignettes used in medical education and found systematic differences in how men and women were presented. In vignettes, men tended to be presented as having more authority, and women as being more caring, including a particularly relevant pattern where doctors in the vignettes were more often male, and nurses more often female.[22]

An authoritative review of the methods used to select junior doctors into specialist training programmes found that those methods were both highly subjective and heavily dependent upon academic results.[23] US medical educators have argued that selecting candidates for medical school and specialist training based on academic results is inherently unjust because it advantages groups that are already over-represented in medical education and practice, and who share the background, experiences, and assumptions of the senior academics who create medical curricula.[24]

The lack of a neutral map of medical knowledge makes it difficult to identify the biases of the individual senior doctors who create medical curricula. While there is no doubt that the published

peer reviewed medical literature contains its own biases, it is the most complete written record of medical knowledge. In addition, unlike the implicit biases that influence individual experts, the biases in the peer reviewed literature can be systematically studied and addressed.[25] Therefore, a map derived from the peer-reviewed medical literature may be the most effective measure by which to identify and address expert bias.

1.4 - Making the structures of scientific knowledge intelligible to human understanding with machine learning

Rapid improvements in computational hardware, ML algorithms, and the collection, organisation, and accessibility of information over the last two decades have made it possible to extract and visualise patterns of knowledge well beyond the comprehension of individual experts in their raw forms. ML algorithms can infer meaningful patterns from any type and volume of data and can be designed to identify known patterns or to search for previously unknown relationships, but they require human oversight and interpretation to be useful.[4]

ML applies rapid mathematical transformations to large sets of data. Some types of human cognition, visual comprehension for example, are similar to ML techniques such as recurrent neural networks, as they involve rapid processing of enormous volumes of data using parallel pathways in the nervous system and brain. However, most cognitive processes relevant to acquiring medical knowledge are slow and linear. For example, learning to formulate a medical diagnosis based on a history and examination involves a sequence of questions and physical investigations testing an evolving series of hypotheses about the reasons for a patient's symptoms.[26] A major consequence of the differences between human and machine information processing is that it is often not possible for humans to understand how ML algorithms work.[27]

The difficulty that humans face in understanding ML algorithms stimulated the field of explainable artificial intelligence (xAI). xAI seeks to develop algorithms in ways that make the internal processes more perceptible to humans, or to provide secondary algorithms that translate the mechanisms that

make them effective into human-understandable form. Algorithms which work in ways that are readily intelligible to humans are technically known as *explainable*, while algorithms that can be understood after translation are described as *interpretable*.[27] There is evidence that doctors are particularly reluctant to rely upon ML results if they do not understand the processes by which the results were obtained whether directly or after interpretation.[28]

The science of science (SoS) is a field designed to understand and describe the structure and evolution of science.[29] As it is impossible for individual human beings to read, understand, and remember even a small fraction of the sum total of scientific (or even just medical) knowledge, the first challenge for SoS analyses is to select from and summarise that knowledge in human understandable ways. Compared with the slow linear presentation of information in language or numbers, visualisation of knowledge abstractions in graphical maps leverages the capacity of human visual systems to rapidly decode in parallel dense knowledge networks based on signals such as distance, colour, and shape.[30]

SoS has generated maps from various data sources, including the relationships between individual researchers, institutions such as universities, registered claims of novel knowledge such as patents, and others.[4] However, of the potential sources, the research and opinion collected in peer-reviewed medical journals is the most authoritative and comprehensive store of accessible knowledge relevant to learning and applying the KSA of medicine.

1.5 - Analysing the peer reviewed medical literature with self-organising maps

At the time of writing the Medline database records information about 33 million articles published in the peer-reviewed medical literature. While the database was first established in 1975, it includes information about articles dating back to the 19th century and is adding more than a million articles every year.[31] Every article is annotated by a set of Medical Subject Headings (MeSH) which indicate the knowledge contained in the article. Annotation is a technical term which means to apply a label with a precise scientific meaning to a specific type of entity. The US-based National Library of

Medicine (NLM) maintains a controlled vocabulary of MeSH with defined meanings which they use to describe all Medline-indexed articles. ML techniques including self-organising maps (SOMs),[30] recurrent neural networks (RNNs),[32] and various forms of network and cluster analysis[4] have been used to produce maps of the knowledge contained within the Medline database.

While Medline is limited to published peer-reviewed medical literature and is free for public use, there are two large commercial alternatives which index a larger set of medical literature alongside scientific publications and other materials including patents (Web of Science and Scopus). Both alternatives provide free access to small subsets of data but require commercial arrangements for full access. In addition, neither alternative provides a controlled vocabulary with the level of detail of Medline's MeSH. As a result, this thesis analyses Medline data rather than the alternatives.

The SOM is a form of artificial neural network loosely modelled on fields of neurons in the human brain.[33] In its standard form, a SOM can be visualised as a square or hexagonal field of individual units called nodes, each one like an individual neuron in a section of the brain involved in visual or auditory perception. While the activity of neurons is determined by the interaction between neuronal properties and the sum of chemical and electrical activity stimulated by other neurons, each SOM node only receives external inputs.

The function of a SOM is similar to the function of a layer of neurons in the brain which receive literal information about the physical properties of the outside world and transform it into representations of the meaning of those physical properties.[33,34] For example, when hearing speech, the auditory cortex receives information about vibrations in the atmosphere detected by cochlear cells in the ear and transforms it into symbolic information about letters and words. When reading text, the visual cortex receives information about light particles entering the retina and similarly transforms it into symbolic information about letters and words.[35]

A SOM is commonly designed as a 2-dimensional grid of nodes, in which each node is an individual processing unit modelled on a neuron. SOMs operate in two distinct modes: a learning mode during

which a large set of stimuli are presented to the SOM multiple times and the nodes learn to process the information contained in those stimuli to match them with abstract categories; and a recognition mode, where individual stimuli are presented and the SOM categorises them based on previous learning. No learning occurs in the recognition phase.

The learning and recognition modes are described in more detail later in the thesis, but the representation of knowledge described by a SOM can be understood with reference to Figure 1-1. A SOM can be trained on any information that can be represented as a vector of numbers. The fictional example in Figure 1-1 shows a 2-dimensional grid of eleven columns and twelve rows of nodes in a self-organising map after they have been trained on a set of inputs that represents the presence or absence of signs and symptoms of disease in individual patients as a vector of 1s and 0s respectively.

After the training phase is complete, each node in a SOM represents a concept related to the training inputs. In the case of Figure 1-1, each node represents a combination of signs and symptoms characteristic of a particular diagnosis or syndrome of disease. During the recognition phase, new patients can be presented to this SOM, represented by the presence or absence of the same signs and symptoms used in the training phase. The SOM will recognise which node is most like the new patients' signs and symptoms and thereby map them to a particular location in the 2 dimensional knowledge space.

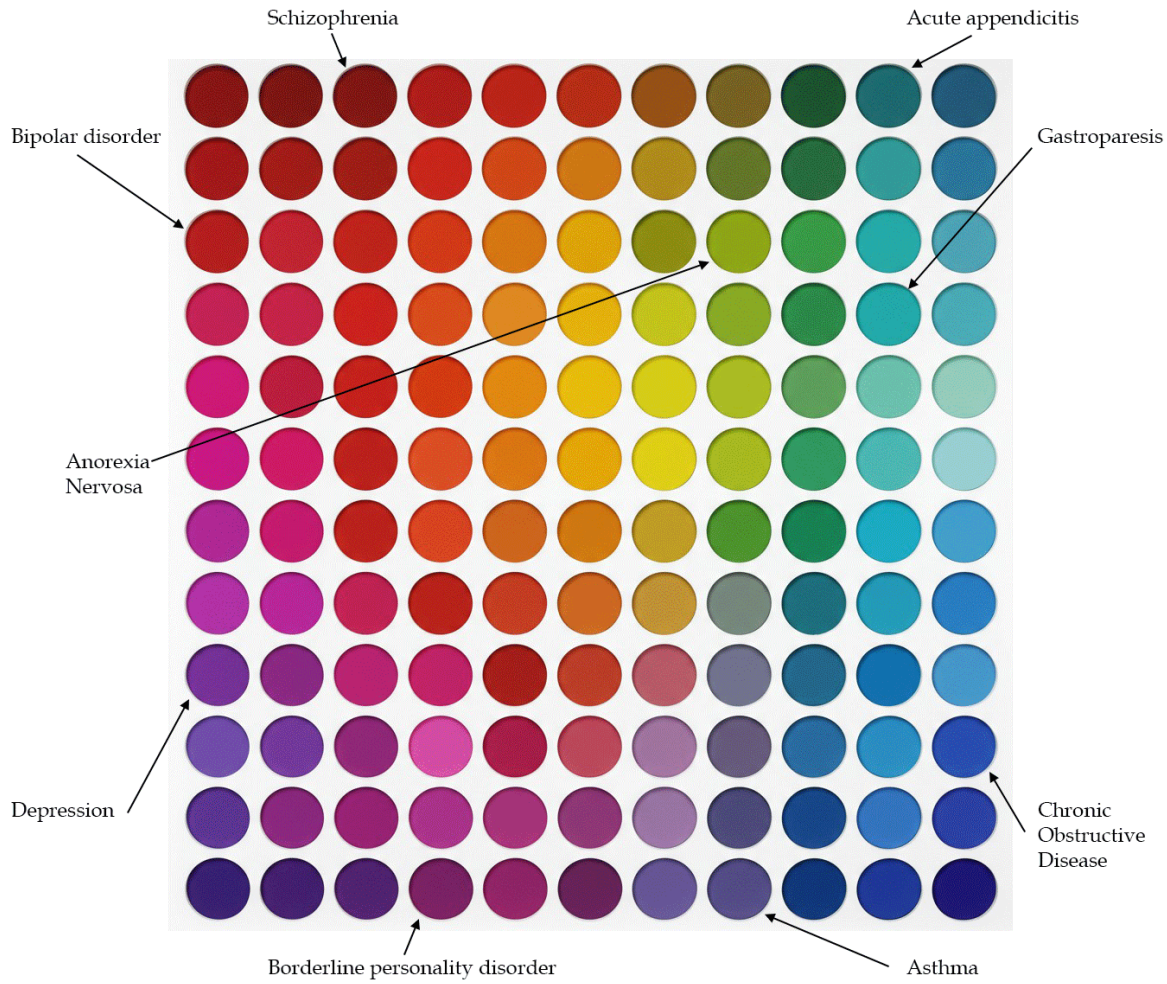


Figure 1-1: Schematic outline of the nodes of a self-organising map trained to recognise patients represented by signs and symptoms

As can be seen, after training, the two dimensions of the SOM represent the knowledge space of possible medical (including psychiatric) diagnoses. Psychiatric diagnoses which are similar because they involve psychosis, including schizophrenia and bipolar disorder, are grouped together at the top left of the SOM. Other psychiatric diagnoses which do not involve psychosis are grouped together at the bottom left, including depression and borderline personality disorder. Finally, medical diagnoses are grouped into gastrointestinal clusters at the top right, and respiratory at the bottom right. Anorexia nervosa, a psychiatric diagnosis which involves significant physical morbidity, particularly

involving the gastrointestinal system, lies between the psychiatric diagnoses on the left and the physical diagnoses on the right, somewhat closer to the gastrointestinal cluster at the top.

1.6 - Validating a self-organising map trained on the Medline database

It is relatively simple to understand the small 2-dimensional (fictional) knowledge space represented by Figure 1-1 in terms of two axes from psychiatric to physical illness in one dimension, and from psychotic to non-psychotic illness and from gastrointestinal to respiratory illness in the other dimension. A 2-dimensional map with tens of thousands of nodes that summarise the 30 million articles, and the close to 30 thousand medical concepts used to label those articles, indexed by the Medline database, is more difficult to comprehend.

One way to confirm that any map derived from the Medline database is internally coherent and externally consistent with real-world knowledge is to use it to analyse the knowledge structures of independently derived knowledge products. Medical textbooks are an ideal candidate for this type of analysis because they are the result of structured selections from the published peer-reviewed medical literature by medical experts, and they include reference lists to support their knowledge claims. Textbooks with multiple editions are particularly useful for this purpose because they can be used to see whether the SOM is consistent with knowledge structures as they evolve over time.

1.7 - Integrating machine learning products with human decision-making: Emerging topics and continuing professional development

To be useful, the results of SOM analysis of Medline data must provide information that can be integrated into the decision-making of medical experts. There is very little research available to guide the integration of ML products into medical education decision-making. The most relevant approach to date compared how a SOM categorised the concepts discussed by an online support group with an alternative algorithm and a human expert.[36] No examples were found that explored the extent to which SOM outputs were useful in decision-making.

An area where ML might contribute to medical education is continuing professional development (CPD), which comprises the formal and informal learning activities completed by medical consultants to maintain or develop their professional competencies and fulfill the requirements of professional registration.[37] In Australia CPD is largely self-directed, with doctors completing a certain minimum number of hours each year engaged in activities they select within parameters defined by regulatory bodies. Reading published peer-reviewed articles is a common CPD activity used to maintain the knowledge, skills, and attributes (KSAs) necessary for competent medical practice.

It is reasonable to assume that the CPD activities selected by doctors might be subject to the same biases that influence the senior clinicians who create and maintain medical training curricula. One means of addressing this form of bias are ML algorithms capable of identifying novel and professionally relevant articles to maintain doctors' KSAs. Ohniwa and colleagues developed an algorithm that used the Medline database to identify emerging MeSH topics in medicine, which they labelled the Increment statistic. Essentially, their algorithm is sensitive to rapid increases in the frequency with which medical topics are being discussed in the peer-reviewed literature.[38,39]

A 2-dimensional SOM trained to describe the knowledge structures and relationships between the MeSH of the Medline database could be used in combination with the Increment statistic to identify regions with an above average density of emerging MeSH topics. This can be understood as conceptually similar to the superimposition of topographic and geographic features like elevation, precipitation, population on a plain political map. Just as colours and shading can show how close the peak of a mountain range is to a city, colours and shading can show how close a MeSH of emerging interest such as Transcranial Magnetic Stimulation (a novel form of psychiatric treatment for depression) is to the diagnosis of Major Depressive Disorder in the knowledge space represented by the published peer reviewed medical literature.

1.8 - Thesis aims and hypotheses

1.8.1 - Aims

Medical curriculum development relies almost entirely upon expert judgement for selecting content, designing assessments, and confirming competency throughout medical training, which introduces the risk of bias. An empirical estimate of the boundaries and relationships between domains of medical knowledge would provide an independent standard to guide these curriculum decisions and identify and reduce bias. ML techniques have recently been developed with the potential to condense and summarise the information contained within massive medical literature databases to perform this function. This thesis aims to:

1. Evaluate the evidence that current methods of selection into specialist medical training are biased against groups which are currently under-represented in specialist training and practice.
2. Apply the self-organising map technique to the Medline database indexing the peer-reviewed medical literature to produce a static map summarising the content and structure of medical knowledge.
3. Extend the self-organising map technique to analyse the structure of knowledge contained in a core psychiatric textbook.
4. Apply bibliometric techniques to the Medline database to demonstrate the capacity to identify emerging topics in the medical literature suitable for use in medical education/professional development.
5. Test the utility of the knowledge produced by ML analysis to guide the continuing professional development of practising psychiatrists.
6. Demonstrate the technical improvements in the self-organising map algorithm required to extend previous analyses from subsets to the entire corpus of the Medline database.

1.8.2 - Hypotheses

To address these aims, this thesis hypothesises that:

1. The methods of selecting junior doctors into specialist training programmes contain systematic biases associated with the under-representation of some groups in specialist training and practice.
2. The structure and limits of the domains of medical knowledge latent in the information contained within large databases indexing the peer-reviewed medical literature can be condensed and summarised in human-intelligible form by self-organising maps.
3. The knowledge represented by the reference lists of a core psychiatric textbook can be interpreted by a self-organising map trained on the Medline database.
4. Topics of emerging interest in the peer-reviewed medical literature can be identified using bibliometric techniques.
5. Practicing psychiatrists will rank articles on topics of emerging interest in the peer-reviewed medical literature identified using bibliometric techniques as more novel and more relevant to practice than articles selected at random.
6. Advances in the hardware and technical improvements in the self-organising map algorithms have been sufficient to integrate all available information from the Medline database in comprehensive maps of medical knowledge.

This thesis describes the testing of these hypotheses across ten chapters, including this general introduction, one chapter explaining the technical and theoretical background of ML models, five manuscripts produced for publication in peer-reviewed journals, one peer reviewed manuscript from the conference proceedings of a presentation at an academic conference, and capped by discussion and conclusion chapters.

1.9 - Methodology

1.9.1 - Research Design

This thesis drew on a variety of ML and bibliometric techniques applied to the large Medline database indexing the peer-reviewed medical literature and a cross-sectional study which compared

the relevance and novelty of articles selected by bibliometric techniques as topics of emerging importance. The ML and bibliometric research projects were designed to demonstrate that it is possible to condense, summarise, and present in human-intelligible form the complete set of information contained in large databases and use this map to understand medical knowledge structures such as the knowledge in a psychiatric textbook. The cross-sectional study was designed to confirm that the information extracted from the databases had meaning and relevance to medical specialists in the context of continuing professional development.

Two ML models based on the SOM were produced. The first ML model extended previous maps examining subsets of the Medline database with a more sophisticated algorithm that allowed for the simultaneous mapping of the entire dataset. The second ML model supported the validity of the first model by demonstrating that it coherently organised the information contained across ten editions of a standard textbook of psychiatry. A bibliometric model was applied to the same Medline database as the SOMs to identify topics of emerging interest. A questionnaire was then offered to practising psychiatrists to evaluate the extent to which these topics of emerging interest were both novel and relevant for the purposes of continuing professional development.

1.9.2 - Key Data Sources and Participants

The bulk of the thesis involved the analysis of the Medline database which indexes the peer-reviewed medical literature, including articles extending back into the 19th Century. One of the ML studies applied the structures derived from the SOM to interpret the structure and relationship of the information contained across ten editions of the core psychiatric textbook *Kaplan and Sadock's Comprehensive Textbook of Psychiatry*.[35]

Participants in the cross-sectional study were practicing psychiatric consultants in Australia and New Zealand who are required to complete annual continuing professional development as a condition of ongoing professional registration. They were recruited through advertisements in emails distributed to Fellows of the Royal Australian and New Zealand College of Psychiatrists (RANZCP). Their

recruitment and involvement were approved by Ethics committees at both James Cook University and at the RANZCP.

1.9.3 - Questionnaire Design

As it was the first of its kind, the questionnaire presented to practising psychiatrists was of necessity designed de novo. As a pilot study the design was intentionally made as simple as possible, focused on the smallest irreducible unit of psychiatric knowledge represented in the ML and bibliometric studies, the title and abstract of published peer-reviewed articles. The questionnaire simply presented a set of articles about psychiatry published in 2022 and selected by the bibliometric method as addressing emerging topics, interspersed among a set of titles and abstracts selected at random from articles published about psychiatry in the same year, and asked a group of psychiatrists recruited through the Royal Australian and New Zealand College of Psychiatrists' emailed newsletters to rank those abstracts in terms of relevance and novelty.

1.9.4 - Interpretation of Results

Results of the ML research were interpreted in terms of their technical properties and their epistemological implications. The technical improvements in hardware and algorithmic advances required to complete the analyses were demonstrated in a technical paper that compared the feasibility and absolute capacity for including the complete set of Medline articles and all MeSH in training SOMs using previous algorithms to the improved algorithms developed during this research. The epistemological qualities and validity of the maps describing the complete set of information contained in the Medline database were explored by demonstrating that they could be used to coherently organise the information contained in the reference lists of the chapters of a core psychiatric textbook.

Results of the bibliometric research on emerging topics were validated by asking a group of practising psychiatrists to rank their ability to select articles that were both novel and relevant for the purposes of continuing professional development.

1.10 - Expected outcomes

As the techniques and experiment in this research apply relatively new techniques in a novel way to a large and complex dataset their expected outcomes should be understood as the first steps in the process of using ML and artificial intelligence to condense and summarise the information contained in large scientific databases to guide medical curriculum development and similar activities.

The static SOM described here is the first to integrate the entire set of information contained in the Medline database in a single map. Its extension was expected to demonstrate that a SOM derived from the entire Medline database can produce a coherent map of the knowledge contained within a psychiatric textbook, and therefore is likely to be able to coherently summarise other documents of specialist medical knowledge; and that it is possible to use a SOM to understand the evolution of structures and relationships of domains of medical knowledge over time.

The emerging topics research demonstrated that it is possible to usefully identify from the published peer-reviewed literature which topics and which articles contain relevant and novel information which may be suitable for inclusion in the process of curriculum development. As the emerging topics algorithm is applied to the same Medical Subject Headings that constitute the concepts mapped by the SOM they serve as an example of the type of information it would be possible to overlay on the SOM.

Finally, assuming that the ML research is successful in achieving its goals, the technical analyses of the methods used to implement the models is expected to expand the boundaries of what is currently possible in terms of amount of data and time/computing hardware required; and to point to areas of potential improvement to further expand those possibilities. An area of particular interest is the possibility of extending the SOM to continuously model the evolution of knowledge over time, which would require further technical and hardware improvements to be feasible.

1.11 - Thesis structure

This thesis was designed to examine the extent to which medical and psychiatric knowledge structures implicit in peer-reviewed research databases such as Medline can be extracted, summarised, and presented in forms useful for human decision-making. The literature review in Chapter 2 defines the problem domain and surveys the ML algorithms commonly used to address such problems. The second part of the chapter describes the epistemological qualities of machine learning algorithms and discusses their validation.

Chapters 3-8 reproduce the submitted manuscripts, and Chapter 9 discusses the implications of the reported research for developing an epistemology that considers differences in the meaning of ML outputs within the algorithms that produce them and to their human users. Chapter 10 draws conclusions regarding the meaning and validity of the MedSOM and suggests future research to integrate ML visualisation into the curriculum development process managed by human experts. The necessity of bridging this machine-human divide is discussed with reference to: the need to reduce existing systematic biases in medical training/education (Chapter 3) using ML techniques that map the peer reviewed medical literature as a unified corpus (Chapter 4); and the identification of emerging topics in that literature (Chapter 5). Chapter 8 examines whether the topics of emerging importance identified by ML algorithms in Chapter 5 match the judgements of a group of senior psychiatrists selecting continuing medical education material. Chapter 6 illustrates technical improvements that allow for the analysis of the entire Medline database with all MeSH rather than the subsets analysed in previous research, and chapter 7 illustrates the validation of the information contained within the MedSOM using the expert-derived knowledge structure represented by the reference lists of a psychiatric textbook.

Table 1-1: Summary of Thesis Outline with Publication Status

Chapter Title	Chapter Contents	Author Contributions	Submission Status
Chapter 1 – General Introduction	This chapter outlined the justification and focus of the current study. Thesis aims, research questions, research design and methodology justification as well as population /study setting, were also discussed.	AA wrote the introductory chapter with JL, TSG and BSM-A reviewing each draft before approving the final version.	N/A
Chapter 2a: - Review of the technical features of ML algorithms for the mapping of scientific knowledge	This chapter reviewed the available literature on alternative ML algorithms for the mapping of scientific knowledge. It summarised the advantages and disadvantages of the alternatives for the specific task of mapping the domains of	AA wrote the review chapter with JL, TSG and BSM-A reviewing each draft before approving the final version.	N/A

	medical knowledge latent in the Medline database, including the relative feasibility of each technique.		
Chapter 2b: Review of the epistemological qualities of ML algorithms for the mapping of scientific knowledge	This chapter reviewed the literature on the range of meaningful information produced by alternative ML algorithms, alongside the literature on explainable artificial intelligence, which describes the features of the products of AI which determine how well they are understood by human experts.	AA wrote the review chapter with JL, TSG and BSM-A reviewing each draft before approving the final version.	N/A
Chapter 3: Systematic review of specialist selection methods with	This chapter reported a systematic review of the range of methods used to select junior doctors into	AA and BSM-A developed the systematic review protocol. AA, BSM-A, and TSG implemented the	Published in *BMC Medical Education* (2021).[40]

implications for diversity in the medical workforce	specialist training programmes focused on whether biases against under-represented groups had been associated with specific techniques.	review. AA developed the first draft of the manuscript. BSM-A, JL, and TSG reviewed and edited the manuscript. All authors read and approved the final manuscript for submission.	
Chapter 4 (Conference Proceedings): Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps	This chapter reported the design and implementation of a SOM trained on the entire corpus of the Medline database and illustrated by differentiating between the Medical and Psychiatric domains of knowledge represented by the map.	AA developed and implemented the SOM. AA developed the first draft of the manuscript. BSM-A, JL, and TSG reviewed and edited the manuscript. All authors read and approved the final manuscript for submission. AA presented the manuscript at MEDINFO 2023 on July 11th, 2023.	Published in *Studies in Health Technology and Informatics*, 2024.[41]

Chapter 5 (Modelling study): Identifying emerging topics in the peer-reviewed literature to facilitate curriculum renewal and development	This chapter described the application of a bibliometric technique to the Medline database to identify emerging topics in the medical literature since the 1970s.	AA developed and implemented the bibliometric modelling. AA developed the first draft of the manuscript. BSM-A, JL, and TSG reviewed and edited the manuscript. All authors read and approved the final manuscript for submission.	Published in *Current Psychology* (2022).[42]
Chapter 6 (Modelling study): Expanding the scope of maps of scientific knowledge with improvements in the self-organising map algorithm	This chapter describes the technical advances required to allow the SOM technique to integrate the entire Medline database and extend it with a temporal function	AA developed and implemented the protocol for analysing the technical features of the SOMs. AA developed the first draft of the manuscript. BSM-A, JL, and TSG reviewed and edited the manuscript. All authors read and approved the final manuscript for submission.	Submitted to *Computing* (27.06.25).[43]

Chapter 7 (Modelling study): Validating the knowledge represented by a self-organizing map with an expert-derived knowledge structure	This chapter described the application of a SOM trained on the Medline database to the reference lists of a core psychiatric textbook. It reported that the SOM provided a coherent and intelligible interpretation of the knowledge domains represented by the reference lists.	AA developed and implemented the application of the SOM to the reference lists. AA developed the first draft of the manuscript. BSM-A, JL, and TSG reviewed and edited the manuscript. All authors read and approved the final manuscript for submission.	Published in *BMC Medical Education* in 2023.[44]
Chapter 8 (Cross-sectional study): Validating the output of a bibliometric method of identifying emerging topics of medical knowledge with practicing psychiatrists	This chapter reported a cross-sectional study which presented practicing psychiatrists recruited from the RANZCP's monthly newsletter with a questionnaire and tested whether the emerging topics identified by bibliometric	AA and BSM-A developed the study protocol. AA implemented the protocol in liaison with the RANZCP. AA produced the first draft of the manuscript. BSM-A, JL, and TSG reviewed and edited the manuscript. All authors read and	Submitted to *Education Journal* (27.06.25).[45]

	techniques could be used to identify novel and relevant articles for continuing professional development.	approved the final manuscript for submission.	
Chapter 9: General Discussion	Discusses the findings of the research and their implications for the use of ML algorithms to condense and summarise large databases of scientific information into maps useful for guiding the decision-making of medical experts. Considers both the specific domain of medical curriculum development but also the potential for other similar intellectual functions.	AA wrote the discussion chapter with JL, TSG and BSM-A reviewing each draft before approving the final version.	N/A

Chapter 10: Conclusions and Recommendations	The conclusions of the research were summarised and formulated into recommendations for extending the techniques covered by the thesis to better realise the promise of maps of scientific knowledge to support complex decision-making.	AA wrote the concluding chapter with JL, TSG and BSM-A reviewing each draft before approving the final version.	N/A

1.12 - References

1. Harden RM. AMEE Guide No. 21: Curriculum mapping: A tool for transparent and authentic teaching and learning. Med Teach. 2001;23(2):123–37.

2. Thomas P, Kern DE, Hughes MT, Chen BY, editors. Curriculum Development for Medical Education: A Six-Step Approach. Third. Baltimore: Johns Hopkins University Press; 2015.

3. The Lancet. Cardiology's problem women. The Lancet [Internet]. 2019;393(10175):959. Available from: http://dx.doi.org/10.1016/S0140-6736(19)30510-0

4. Boyack KW, Klavans R. Creation and Analysis of Large-Scale Bibliometric Networks. In: Glänzel W, Moed HF, Schmoch U, Thelwall M, editors. Springer Handbook of Science and Technology Indicators. Cham, Switzerland: Springer Nature; 2019. p. 187–212.

5. Kruse A, Schmitt E. Education in Psychiatry.

6. Tonelli MR. The philosophical limits of evidence-based medicine. Vol. 73, Academic Medicine. 1998. p. 1234–40.

7. Hean S, Green C, Anderson E, Morris D, John C, Pitt R, et al. The contribution of theory to the design, delivery, and evaluation of interprofessional curricula: BEME Guide No. 49. Med Teach. 2018 Jun 3;40(6):542–58.

8. Tonelli MR, Bluhm R. Teaching Medical Epistemology within an Evidence-Based Medicine Curriculum. Teach Learn Med. 2020;33(1):98–105.

9. Harden RM, Grant J, Buckley G, Hart IR. BEME Guide No 1: Best Evidence Medical Education. Med Teach. 1999;21(6):553–62.

10. Cook DA, Reed DA. Appraising the Quality of Medical Education Research Methods: The Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Academic Medicine. 2015;90(8):1067–76.

11. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49:560–75.

12. Ibrahim H, Juve AM, Amin A, Railey K, Andolsek KM. Expanding the Study of Bias in Medical Education Assessment. Vol. 15, Journal of Graduate Medical Education. Accreditation Council for Graduate Medical Education; 2023. p. 623–6.

13. Van Spall HGC, Lala A, Deering TF, Casadei B, Zannad F, Kaul P, et al. Ending Gender Inequality in Cardiovascular Clinical Trial Leadership: JACC Review Topic of the Week. Vol. 77, Journal of the American College of Cardiology. Elsevier Inc.; 2021. p. 2960–72.

14. D'Eon M, Crawford R. The elusive content of the medical-school curriculum: A method to the madness. Med Teach. 2005 Dec;27(8):699–703.

15. Densen P. Challenges and Opportunities Facing Medical Education. Trans Am Clin Climatol Assoc. 2011;122:48–58.

16. Carr PL, Raj A, Kaplan SE, Terrin N, Breeze JL, Freund KM. Gender differences in academic medicine: Retention, rank, and leadership comparisons from the national faculty survey. Vol. 93, Academic Medicine. Lippincott Williams and Wilkins; 2018. p. 1694–9.

17. Verdonk P, Benschop YWM, De Haes HCJM, Lagro-Janssen TLM. From gender bias to gender awareness in medical education. Advances in Health Sciences Education. 2009 Mar;14(1):135–52.

18. Malina D, Amutah C, Greenidge K, Mante A, Munyikwa M, Surya SL, et al. Misrepresenting Race-The Role of Medical Schools in Propagating Physician Bias. New England Journal of Medicine. 2021;384(9):872–8.

19. Kamran SC, Winkfield KM, Reede JY, Vapiwala N. Intersectional Analysis of U.S. Medical Faculty Diversity over Four Decades. New England Journal of Medicine. 2022;386(14):1363–71.

20. Chadwick AJ, Baruah R. Gender disparity and implicit gender bias amongst doctors in intensive care medicine: A 'disease' we need to recognise and treat. J Intensive Care Soc. 2020 Feb 1;21(1):12–7.

21. Teherani A, Hauer KE, Fernandez A, King TE, Lucey C. How small differences in assessed clinical performance amplify to large differences in grades and awards: A cascade with serious consequences for students underrepresented in medicine. Vol. 93, Academic Medicine. Lippincott Williams and Wilkins; 2018. p. 1286–92.

22. Arsever S, Broers B, Cerutti B, Wiesner J, Dao MD. A gender biased hidden curriculum of clinical vignettes in undergraduate medical training. Patient Educ Couns. 2023 Nov 1;116.

23. Roberts C, Khanna P, Rigby L, Bartle E, Llewellyn A, Gustavs J, et al. Utility of selection methods for specialist medical training: A BEME (best evidence medical education) systematic review: BEME guide no. 45. Med Teach. 2018;40(1):3–19.

24. Christophers B, Marr MC, Pendergrast TR. Medical School Admission Policies Disadvantage Low-Income Applicants. Permanente Journal. 2022;26(2):172–6.

25. Bradley SH, DeVito NJ, Lloyd KE, Richards GC, Rombey T, Wayant C, et al. Reducing bias and improving transparency in medical research: a critical overview of the problems, progress and suggested next steps. Vol. 113, Journal of the Royal Society of Medicine. SAGE Publications Ltd; 2020. p. 433–43.

26. Balogh EP, Miller BT, Ball JR. The Diagnostic Process. In: Improving Diagnosis in Health Care. Washington, DC: National Academies Press; 2016. p. 1–472.

27. Silva A, Schrum M, Hedlund-Botti E, Gopalan N, Gombolay M. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. Int J Hum Comput Interact. 2022;39(7):1390–404.

28. He X, Hong Y, Zheng X, Zhang Y. What Are the Users' Needs? Design of a User-Centered Explainable Artificial Intelligence Diagnostic System. Int J Hum Comput Interact. 2022;39(7):1519–42.

29. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, et al. Science of science. Science (Journal). 2018;359(6379).

30. Skupin A, Biberstine JR, Börner K. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS One. 2013;8(3).

31. Kastrin A, Hristovski D. Disentangling the evolution of MEDLINE bibliographic database: A complex network perspective. J Biomed Inform [Internet]. 2019;89(June 2018):101–13. Available from: https://doi.org/10.1016/j.jbi.2018.11.014

32. Zitt M, Lelu A, Cadot M, Cabanac G. Bibliometric Delineation of Scientific Fields. In: Glänzel W, Moed HF, Schmoch U, Thelwall M, editors. Springer Handbook of Science and Technology Indicators. Cham, Switzerland: Springer Nature; 2019. p. 25–68.

33. Kohonen T. Self-organizing maps. 3rd ed. Berlin: Springer; 2001.

34. Kohonen T. Physiological Interpretation of the Self-Organizing Map Algorithm. Neural Networks. 1993;6(7):895–905.

35. Sadock BJ, Sadock VA, Ruiz P, editors. Kaplan and Sadock's Comprehensive Textbook of Psychiatry. 10th ed. New York: Wolters Kluwer; 2017.

36. Orwig RE, Chen H, Nunamaker JF. A Graphical, Self-Organizing Approach to Classifying Electronic Meeting Output. Journal of the American Society for Information Science. 1997;48(2).

37. AHPRA. Continuing Professional Development [Internet]. AHPRA Website. 2025 [cited 2025 Jun 25]. Available from: https://www.ahpra.gov.au/Registration/Registration-Standards/CPD.aspx

38. Ohniwa RL, Hibino A. Generating process of emerging topics in the life sciences. Scientometrics [Internet]. 2019;121(3):1549–61. Available from: https://doi.org/10.1007/s11192-019-03248-z

39. Ohniwa RL, Hibino A, Takeyasu K. Trends in research foci in life science fields over the last 30 years monitored by emerging topics. Scientometrics. 2010;85(1):111–27.

40. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Systematic review of specialist selection methods with implications for diversity in the medical workforce. BMC Med Educ. 2021 Dec 1;21(1).

41. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli B. Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps. Stud Health Technol Inform [Internet]. 2024;310:795–9. Available from: https://raw.githubusercontent.com/AndrewAmosJCU/PsychSOM/main/ColorCoded.png

42. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Identifying emerging topics in the peer-reviewed literature to facilitate curriculum renewal and development. Current Psychology [Internet]. 2022 Dec 12;42:30813–24. Available from: https://link.springer.com/10.1007/s12144-022-04090-y

43. Amos A, Lee K, Gupta Sen T, Malau-Aduli BS. Novel sparse matrix algorithm expands the feasible size of a self-organizing map of the knowledge indexed by a database of peer-reviewed medical literature. 2025;Submitted. Available from: https://orcid.org/0000-

44. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Validating the knowledge represented by a self-organizing map with an expert-derived knowledge structure. BMC Med Educ. 2024;24(416).

45. Amos AJ, Lee J, Sen Gupta T, Malau-Aduli BS. Testing the validity of emerging topics in psychiatry identified by a machine learning algorithm with a sample of working psychiatrists. Comput Biol Med. 2025;Submitted.

Chapter 2: Bibliometric application of machine learning algorithms for the visualisation of science

This thesis relies on concepts and techniques from the very distinct domains of medical education, computer science, and information science. While the published and submitted chapters provide sufficient background to understand the medical education concepts and techniques, some of the necessary background to the computer science and information science is assumed knowledge in expert journals and therefore not completely described in those manuscripts.

This chapter provides the necessary background about computer science and information science topics for medical education experts to understand the published and submitted materials.

2.1 - Scientometrics and bibliometrics

Science comprises a set of processes for investigating the reliability of hypotheses about existence, implemented by individuals organised in networks, resulting in diverse products including written documents such as research articles, structured and unstructured datasets, and legal or administrative instruments like patents and clinical guidelines.[1] Scientometrics is the quantitative study of the processes, networks, and products of science focused on understanding the development of scientific methods, the accumulation of scientific knowledge, and the social and material impact of science.[2]

Bibliometrics is the branch of scientometrics which examines the written products of science including journals, articles, citations, and grey literature such as patents. In the 20th century, the development of databases which systematically record basic information about science products revolutionised scientific practice in many ways.[3] For the first time it was feasible to comprehensively survey and evaluate all knowledge in defined areas of scientific endeavour and trace the evolution of knowledge domains and specific topics over time.[2]

Modern bibliometrics started in the 1970s as increased access to reference lists along with accelerating computing power combined to allow citation analyses which could identify the most

important research papers in each field, and their influence on later research. Rapidly declining costs and equally rapidly improving performance of digital storage and computation allowed scientific databases to include ever more detailed descriptions of scientific products, to the point where it is technically possible for all publicly available scientific products to be more or less instantly retrieved and analysed, including text, graphics, and other modalities such as audio.[2]

2.2 - Bibliometric analysis

Bibliometrics comprises techniques that measure and analyse qualities of given sets of written science products.[2] Bibliometrics has focused on two main tasks which share analytic techniques but serve different goals. The more commercial task is the evaluation of scientific merit with reference to some specific metric such as productivity, influence, or innovation. The task more relevant to the current thesis is the visualisation of implicit scientific patterns, or mapping the knowledge discovered using scientific methods, by analysing written science products.

The two main analytic techniques of bibliometrics are citation analysis, which analyses references embedded in written science products to describe relationships between written science products; and content analysis, which analyses features used within written science products, such as the pattern of phrases used in the text, to make inferences about the meaning of those products. The most common type of content analysis is word analysis.[2]

In its early days, citation analysis focused on identifying the most important articles in each knowledge domain and their influence on later articles. Later it became possible to use citation analysis to make more sophisticated inferences, such as evaluating the productivity of individual researchers or research units such as universities. Co-citation analysis was a prominent early form, which measured the similarity between articles based on the pattern of shared and different references they cited.[2]

Figure 2-1 provides an example of the visual output of a co-citation analysis. The figure represents individual journals as circular nodes, and represents the relationship between journals as lines, with

the thickness of the lines dependent upon the number of citations of articles in one journal made in articles of the other journal. The total number of citations of articles in each journal is indexed by the size of the font used to write the name of the journal and the radius of its node. Finally, journals are collected into four clusters indicated by the colours green, red, blue, and yellow, with the similarity of individual pairs of journals indicated by the distance between them across the two dimensions of the map.

Figure 2-1 illustrates how much information about the structure of scientific knowledge contained within a set of related medical journals can be condensed into a 2-dimensional map. It is immediately apparent that there are four clusters of journals, that the most cited journals across clusters are *Clinical Orthopedics and Related Research* and *Bone Joint Journal*, and that the most idiosyncratic journals are those in two remote clusters at the top and top left of the figure representing radiology and pediatrics respectively. An expert in the field would no doubt see additional meaning in the clusters.

Due to greater storage requirements and computational complexity, content analysis developed later than citation analysis.[2] The earliest form was co-word analysis, similar to co-citation analysis, which measured the similarity between articles based on the number of content-relevant words used by both articles. This required the creation of controlled vocabularies which differentiate domain-specific words from general purpose words. For example, a controlled vocabulary for psychiatry would include the word schizophrenia but exclude general purpose words such as “the”. Co-word analysis would rate two articles that both use the word “schizophrenia” frequently as more similar than two articles that did not both use any words in the controlled vocabulary.

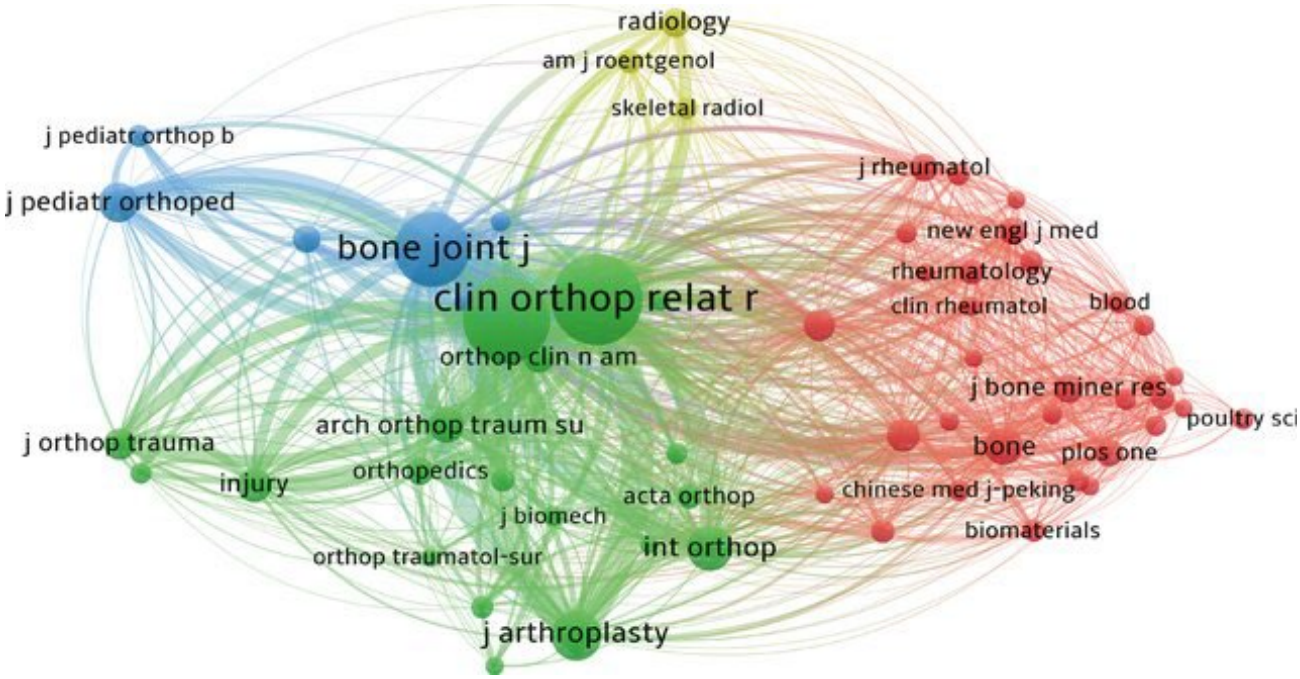


Figure 2-1: Example of co-citation analysis licensed under CC-BY-4.0 (https://creativecommons.org/licenses/by/4.0/)[4]

Figure 2-2 provides an example of the visual output of co-word analysis. Note that meaningful phrases are often treated as indivisible words for the purposes of co-word analysis, as in this case. While Figure 2-1 represents the number of times articles in journals cite each other, Figure 2-2 represents the number of times words with specific meanings co-occur in different articles. It has similar structural features to the co-citation analysis, with large node radius and font size indicating greater frequency of a phrase, the size of the lines between two phrases representing the number of times the phrases co-occur in different articles, and the colours representing meaning clusters. It is common for clusters in co-word analyses to overlap more than clusters in co-citation analyses, for reasons that are not germane to this example.

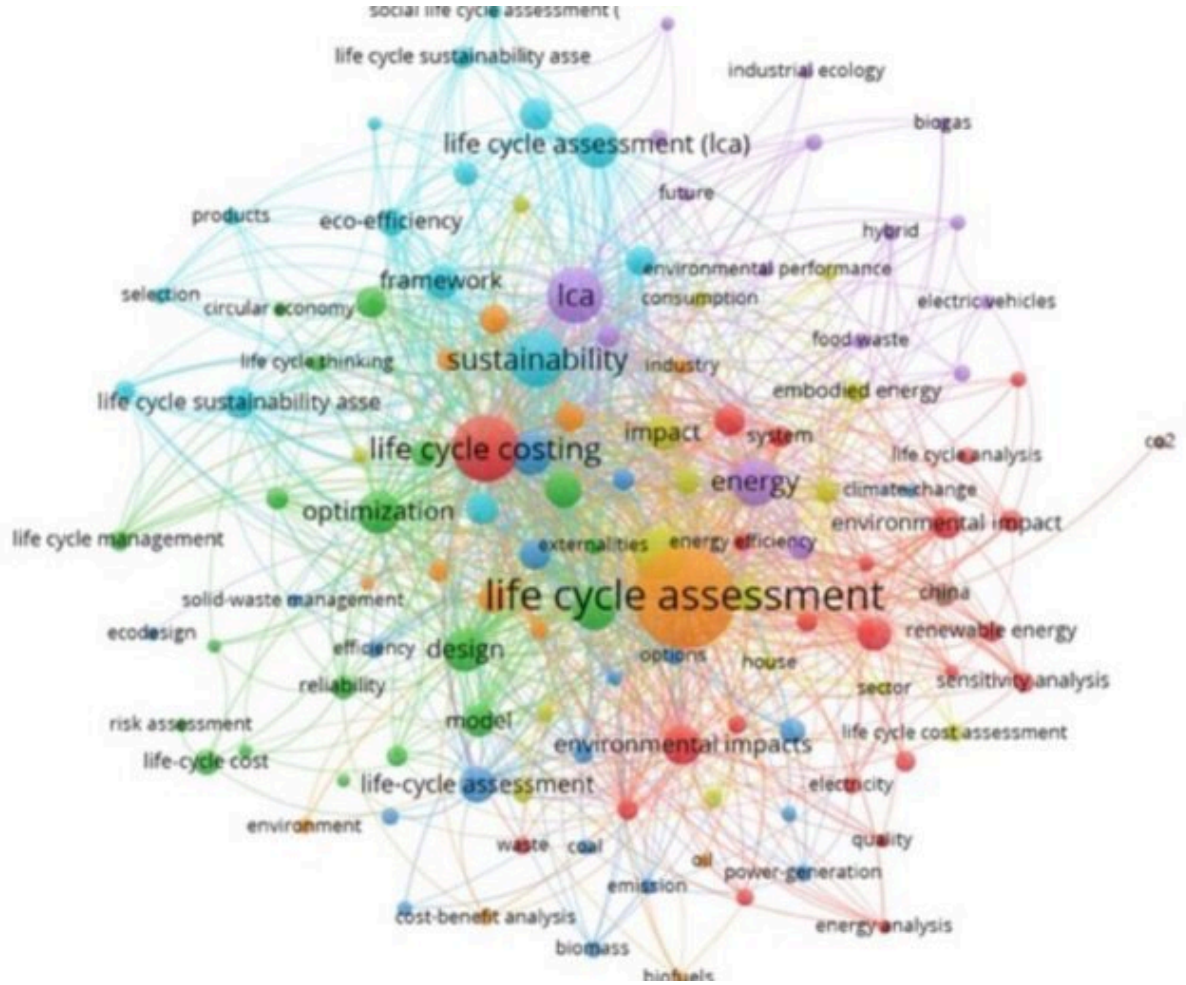


Figure 2-2: Example of co-word analysis licensed under CC-BY-4.0 (https://creativecommons.org/licenses/by/4.0/)[5]

2.3 - Evaluating scientific productivity with bibliometrics

Bibliometric analysis was greatly accelerated by the increased storage and access to scientific data that occurred with the development of the internet. Since 2004 there has been a proliferation of databases which record detailed descriptions of all written science products, such as lists of authors, reference lists, and date of publication. Some cover all branches of science, like Scopus, Web of Science, and Google Scholar, while some cover specific fields, such as Medline, which collects data about all peer-reviewed medical research in journals of sufficient quality and persistence.[3]

The expanded capacity and increasingly commercial nature of bibliometrics in the 21st century has focused attention on improving the productivity and profitability of scientific endeavours. The best-

known bibliometric function is the calculation of the relative merit of different journals. The traditional measure was the journal impact factor (JIF), calculated as the ratio between the number of times that articles in that journal were cited by articles in other journals and the number of articles published in that journal.[6]

In the 21st century increased data access and computational capacity have allowed for the development of other estimates of the relative merit of journals, researchers, and research units such as universities. The Hirsch-index or h-index has been used to measure the merit of an individual researcher's publication history, although it has been criticized for valuing quantity over quality of output.[2] Universities have been ranked based on research quality by bibliometric measures such as the Leiden ranking, as well as hybrid measures that include both research and teaching quality, such as the QS ranking.

2.4 - Identifying patterns implicit in scientific publications with bibliometrics

While there are obvious commercial and administrative uses for estimates of scientific productivity or merit, such as the JIF, bibliometric techniques are increasingly used to reveal previously intangible patterns implicit within the entire corpus of published scientific literature.

For example, Brück used bibliometric methods to analyse gender gaps in the authorship of leading medical journals.[7] It is generally assumed that the first author of a scientific paper was most directly responsible for initiating, implementing, and interpreting the paper's contents, making them the lead author. Likewise, the last author is often assumed to have a less direct but perhaps more senior role, such as leading the laboratory which hosted the research, or providing supervision to a project within which the research was done.

Brück reported that leading authors were twice as likely to be male as female, while senior authors were three times more likely to be male. This is consistent with an overall pattern of a persisting gender gap that is gradually declining but declining least quickly at the most senior levels. There was marked variability between regions, with Asian countries showing much larger gender gaps than

European countries. Brück describes more detailed patterns emerging from the data, including evidence that part of the lower citation count for female-first-authored papers could be explained by less strategic use of keywords by women.[7]

Brück modelled the bibliometric data using two dimensional linear regressions and other low-dimensional statistics, but more powerful methods have recently been developed that allow for the identification of patterns implicit in much more complex datasets. Skupin and colleagues applied the self-organising map (SOM) algorithm developed by Kohonen[8] to the medical research database Medline to detect the patterns implicit in a knowledge space represented by 2300 dimensions and present them in a 2-dimensional map.[9]

2.5 - Applying self-organising maps to bibliometric data

A technical description of SOMs is provided in the published manuscripts that follow these introductory chapters (particularly the addendum to chapter 4 at page **Error! Bookmark not defined.**), but at this point it is useful to understand the purpose of applying the SOM algorithm to bibliometric data by analogy to co-citation and co-word analysis.

Brück[7] and the co-citation and co-word examples above illustrate the use of bibliometric analysis to describe the relationship between well-defined and relatively small sets of meaningful units. Brück analysed the gender gap with reference to less than ten variables using 2-dimensional linear regression and other simple statistics. Co-citation and co-word analyses usually describe the relationships between a few tens or a few hundreds of articles, journals, or words.

In addition, each of the examples analyse known relationships – the linear relationship between a few independent and dependent variables in a regression, or between a small set of articles, journals, or words for co-citation and co-word analyses.

Among the advantages of ML algorithms is that they can apply bibliometric analysis to data of any number of dimensions and detect patterns that were previously unknown. They can analyse highly

structured data, such as the pattern of citations examined by co-citation analysis, or unstructured data, like the text records of posts to social media sites.

An illustration of the power of bibliometric analysis using ML algorithms is provided by Skupin and colleagues. They applied a type of neural network called a self-organising map (SOM) to a subset of the data contained in the medical reference database Medline to produce a visualisation conceptually related to a co-word analysis.[9] A simple example of a SOM was briefly described in the first chapter of the thesis.

At the time of Skupin et al.'s analysis the Medline database recorded information about more than 20 million articles published in peer-reviewed medical journals. Alongside information like article title, authors, and references, Medline employs people to annotate each article with technical signifiers called Medical Subject Headings (MeSH). Medical Subject Headings are an example of a controlled vocabulary – a set of labels that each precisely define one characteristic of the content of a medical article.

A traditional co-word analysis, even of the Medline subset considered by Skupin et al. of 2 million articles and the most common 2,300 MeSH of the 23,347 used at the time, was both computationally impossible to complete, and impossible to visualize. By training a SOM on the data, Skupin et al. were able to produce a 2-dimensional map that is conceptually similar to a traditional co-word analysis. However, instead of representing the similarity between words based on their co-occurrence in article pairs, the SOM learns the most efficient way to represent the similarities between all articles in a 2-dimensional space by iteratively shifting them closer to other articles with which they share relatively many MeSH and away from articles with which they share relatively few MeSH.

Figure 2-3 shows the SOM trained by Skupin et al. on more than 2 million articles published between 2004 and 2008 and indexed by Medline, considering the most common 2,300 MeSH. While similar to the co-word analysis presented in Figure 2-2, the SOM is far more detailed and condenses far more

information than traditional analyses. The central panel in Figure 2-3 can be thought of as a representation of the knowledge space of all articles published in major peer-reviewed medical journals between 2004-2008.

Like Figure 2-2, the SOM in Figure 2-3 describes overlapping clusters of related concepts, although at a much finer level of detail. The top middle panel shows a cluster of major risk factors for heart disease in blue, including aging, body mass index, blood pressure, heart rate, obesity, and exercise. These are laid over less prominent risk factors for the related condition of diabetes, including obesity, blood pressure, blood glucose, insulin, exercise, and weight loss.

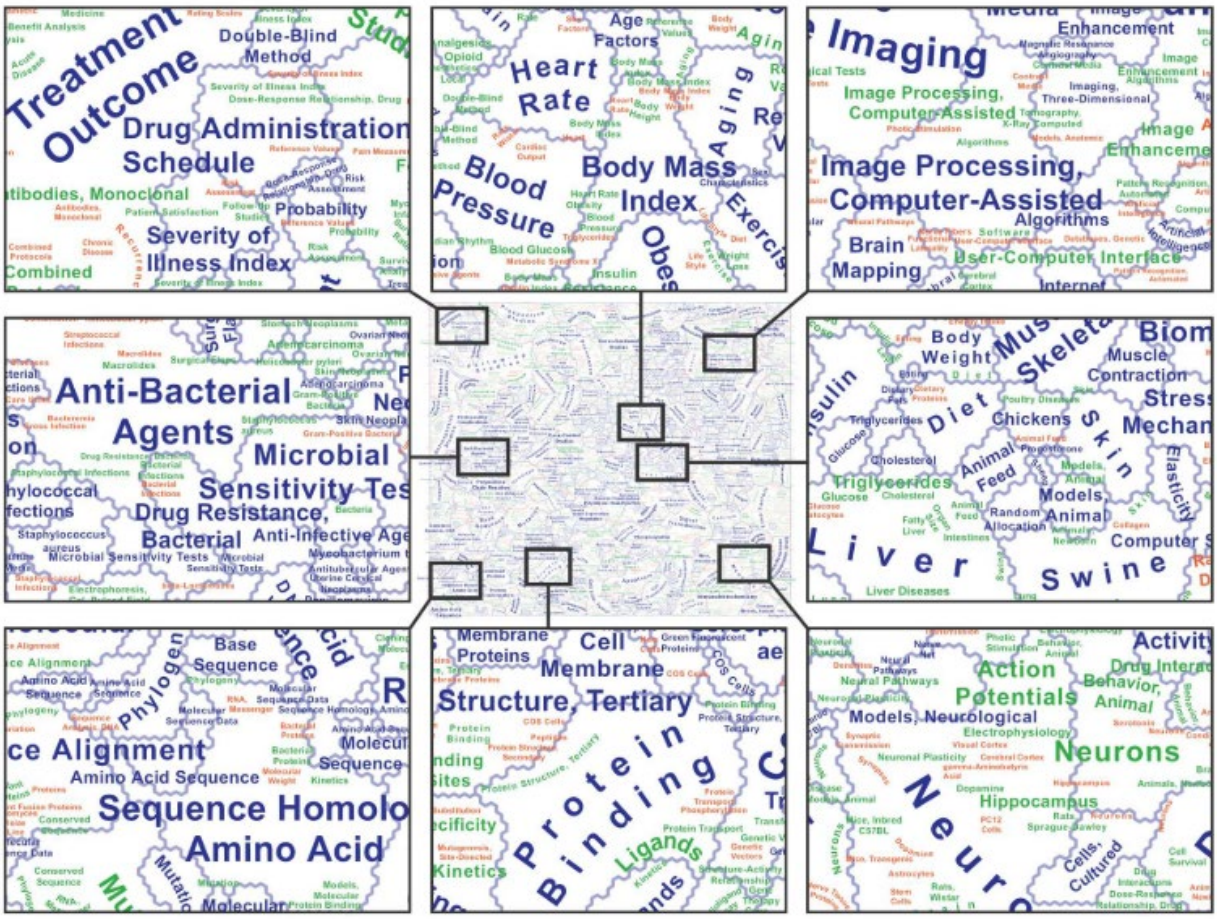


Figure 2-3: Skupin's map of the medical literature showing regions of interest is licensed under CC-BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

2.6 - Promise and pitfalls of self-organising maps for understanding scientific knowledge

Traditional bibliometric techniques like co-citation and co-word analysis are usually readily interpretable. The ability of ML algorithms like SOMs to visualize an enormous set of data in a 2-dimensional map is both an advantage and a disadvantage. While it becomes possible to visually represent an enormous amount of information very concisely, at the same time it becomes more difficult to understand and validate the meaning of the visual summary.

2.7 - Epistemological qualities of machine learning algorithms for mapping scientific knowledge

Like many scientific and technical advances, the application of machine learning algorithms to the problem of understanding the nature and development of scientific knowledge combines opportunities and problems. Among the opportunities are the ability to summarise and visualize features of the entire corpus of articles published in high quality peer-reviewed scientific journals in ways that are intelligible to individual human beings.

The greatest problems associated with ML-based bibliometrics are caused by the reality that the mechanisms by which they detect meaningful patterns are not currently intelligible to human beings. This may be a necessary corollary of their power – for example, it may simply not be possible for human beings to understand the mechanisms by which the SOM described by Skupin et al.[9] summarised the entire corpus of more than 2 million peer-reviewed medical articles between 2004 and 2008 (see Figure 2-3).

The inscrutability of ML algorithms has resulted in the growth of the field of explainable artificial intelligence (xAI). xAI starts with the acknowledgement that we do not understand how or why the most powerful ML algorithms work, and we often do not fully understand the meaning of the results they return. xAI is then designed to address the practical and theoretical problems that result.[10–12]

An important practical problem arising from the inscrutability of ML algorithms is experts' reluctance to rely upon results of processes that they do not understand. A good illustration of the tension

between the promise and problems of ML algorithms is that one of the most powerful early applications of ML algorithms in medicine has been the interpretation of imaging results. ML algorithms have rapidly become able to detect pathological patterns in imaging studies like x-rays and MRIs more accurately and far more quickly than human experts, but human experts have proven reluctant to rely upon those results when they don't understand how the algorithm reached its conclusions.[13]

2.8 - Methodological and epistemological barriers to the use of ML algorithms in medicine

Conceptually, barriers to the use of ML algorithms for medical tasks like the identification of pathology in imaging studies fall into two categories, both caused by the inability to understand how they work. The first is methodological and may be expressed in terms of a question: how can we safely use the results of ML algorithms in medical practice? For example, radiologists who do not understand how a ML algorithm works to detect cancerous lesions on chest x-rays may be reluctant to rely upon them for diagnostic decisions out of fear that their confusion will result in patient harms.[13]

The second is epistemological and may be expressed in terms of a different question: what do the results of ML algorithms mean? As methodological barriers are generally driven by a particular medical task, such as diagnosis, they are usually well-defined, while epistemological barriers are often more difficult to understand.

This can be illustrated by considering the relationship between two of the bibliometric techniques considered earlier in this chapter. The co-word analysis illustrated in Figure 2-2 describes a fairly intelligible set of relationships between around 100 phrases based on their co-occurrence in a few thousand medical publications. The SOM illustrated in Figure 2-3 describes the far more complicated relationships between 2,300 medical subject headings (MeSH) across more than two million peer-reviewed medical articles over the years 2004-2008. It presents overlapping sets of conceptually related phrases forming clusters of meaning that require further interpretation to be useful.

Babushkina & Votsis, among others, have pointed out that given the possibility that the mechanisms that allow ML algorithms to work may never be clearly intelligible to humans, the meaning of products like the SOM in Figure 2-3 may never be fully described. Babushkina & Votsis conclude that as a result "[c]ases like AI assisted diagnosis are no longer bound *merely* by the norms of human epistemology" (p22).[14]

2.9 - AI assisted medicine and the need for a hybrid epistemology

Babushkina & Votsis argue that the case of ML algorithms that more accurately detect cancerous lesions in chest x-rays than humans by means doctors do not understand demonstrates the need for a hybrid epistemology which recognizes the integration of ML generated information into an individual human being's cognitive processing of multiple sources of meaningful information.[14]

Babushkina & Votsis note the tendency to attempt to understand the epistemology of ML products by analogy to human cognition.[14] In the case of AI detection of pathology from imaging studies, this approach attempts to interpret what the AI does as similar to the human expert approach of looking for characteristic sets of deviations from variations of normal based on an understanding of indicators of healthy and pathological human anatomy.

Human doctors who interpret imaging studies consciously integrate a store of explicit knowledge about anatomy learned by study, dissection, and, increasingly, by 3-d representations of gross anatomy overlaid with representations of microscopic and functional anatomy, with a store of knowledge about healthy development and the progression of disease.[15]

ML algorithms may have no store of knowledge independent of the information implicit in the images on which they are trained. While in some cases it has proved possible to infer some of the indicators used by ML algorithms to perform tasks like identifying lesions, there is a high degree of uncertainty for all but the simplest tasks.[16]

As it is undeniable that the most powerful ML algorithms used for medical tasks like image analysis do not, and in most cases can not rely upon the norms of human epistemology, Babushkina & Votsis

argue that "[t]here is a pressing need to analyze the epistemic capabilities of different types of algorithmic solutions, estimate how these capacities relate to the production of knowledge, and deduce normative constraints that apply to them. This is crucial if we are to realistically assess what sort of conclusions we are warranted to draw from AI algorithms" (p22).[14]

2.10 - Validating the meaning of a self-organising map of science

In the context of medical practice, the meaning of information is constrained by the intended use of the information. Considered as part of the assessment of a patient presenting in acute distress to an emergency department, the presence or absence of a risk factor for chronic disease such as high blood pressure has a different meaning when the presenting complaint is chest pain than when it is a broken leg. It has a different meaning again when considered as part of an analysis predicting future health system demands based on the characteristics of adults approaching retirement.

Cartographers use indicators like colour, position, size, and perspective to superimpose overlapping sets of information about disparate domains including geography, elevation, population, and so on. Skupin's SOM describing the medical literature was a proof of concept illustrating that the same indicators could be used to provide a dense array of overlapping domains of medical information in a way that is intelligible to human beings.[9] However, because it was not designed to play a part in the completion of any particular task, its meaning was poorly defined.

Considering a SOM like Skupin's as a potential input to the process of curriculum development helps characterise the types of meanings that it might usefully elicit from the published peer-reviewed medical literature. At the most superficial level, Skupin's SOM describes the relative organisation of medically meaningful phrases (MeSH) across a 2-dimensional representation of all medical knowledge. Due to technical constraints, Skupin's SOM was limited to the articles published between 2004 and 2008 and applied to a subset of the MeSH (2,300 of 23,347 total MeSH).

In the absence of an ultimate standard of meaning against which to measure such a SOM it is necessary to find other comparisons. To demonstrate that a medical SOM (MedSOM) like Skupin's

could be valid for the purposes of constructing a medical curriculum it would be useful to compare it to existing knowledge structures used for the same purpose, and to verify that it generates outputs that are intelligible to medical doctors.

For the purposes of this thesis it was decided to demonstrate that technical and technological improvements have made it possible to construct a MedSOM trained on the entire set of articles indexed by the Medline database; to show that the knowledge structures elicited were consistent with the knowledge structures of a leading psychiatric textbook; that ML algorithms can identify which MeSH are emerging as topics of potential interest for integration into a medical curriculum; and to verify that MeSH identified in this way were perceived as meaningful by practising psychiatrists.

2.11 - **References**

1. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, et al. Science of science. Science (Journal). 2018;359(6379).

2. van Raan A. Measuring Science: Basic Principles and Application of Advanced Bibliometrics. In: Glänzel W, Moed HF, Schmoch U, Thelwall M, editors. Springer Handbook of Science and Technology Indicators. Cham, Switzerland: Springer Nature; 2019. p. 237–80.

3. Zitt M, Lelu A, Cadot M, Cabanac G. Bibliometric Delineation of Scientific Fields. In: Glänzel W, Moed HF, Schmoch U, Thelwall M, editors. Springer Handbook of Science and Technology Indicators. Cham, Switzerland: Springer Nature; 2019. p. 25–68.

4. Wu H, Cheng K, Tong L, Wang Y, Yang W, Sun Z. Knowledge structure and emerging trends on osteonecrosis of the femoral head: a bibliometric and visualized study. J Orthop Surg Res. 2022 Dec 1;17(1).

5. Yi Y, Manzardo A, Lavagnolo MC. Inclusion of prevention activities in LCA and LCC of construction waste management: a review. In: Proceedings of the Conference of Waste Management Procedures

[Internet]. Padua: University of Padua; 2022. Available from: https://www.researchgate.net/publication/362491095

6. Larivière V, Sugimoto CR. The Journal Impact Factor: A Brief History, Critique, and Discussion of Adverse Effects. In: Springer Handbook of Science and Technology Indicators. Berne: Springer Nature; 2019. p. 3–24.

7. Brück O. A bibliometric analysis of the gender gap in the authorship of leading medical journals. Communications Medicine. 2023 Dec 1;3(1).

8. Kohonen T. Self-organizing maps. 3rd ed. Berlin: Springer; 2001.

9. Skupin A, Biberstine JR, Börner K. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS One. 2013;8(3).

10. Silva A, Schrum M, Hedlund-Botti E, Gopalan N, Gombolay M. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. Int J Hum Comput Interact. 2022;39(7):1390–404.

11. Sanneman L, Shah JA. The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. Int J Hum Comput Interact. 2022;38(18–20):1772–88.

12. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. Applied Sciences (Switzerland). 2021 Jun 1;11(11).

13. Najjar R. Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. Diagnostics. 2023 Sep 1;13(2760).

14. Babushkina D, Votsis A. Epistemo-ethical constraints on AI-human decision making for diagnostic purposes. Ethics Inf Technol. 2022 Jun 1;24(2).

15. Láinez Ramos-Bossini AJ, López Cornejo D, Redruello Guerrero P, Ruiz Santiago F. The Educational Impact of Radiology in Anatomy Teaching: A Field Study Using Cross-Sectional Imaging and 3D Printing for the Study of the Spine. Acad Radiol. 2024 Jan 1;31(1):329–37.

16. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Vol. 42, Medical Image Analysis. Elsevier B.V.; 2017. p. 60–88.

Chapter 3: Systematic review of specialist selection methods with implications for diversity in the medical workforce

Authors:

Andrew James Amos[*1], MB.BS, Kyungmi Lee[2], PhD, Tarun Sen Gupta[1], PhD, Bunmi S. Malau-Aduli[1,3], PhD

* Corresponding Author

1 College of Medicine and Dentistry, James Cook University, Townsville, Australia

2 College of Science and Engineering, James Cook University, Cairns, Australia

3 School of Medicine and Public Health, University of Newcastle, Newcastle, Australia

Citation: Amos A, Lee K, Sen Gupta T, Malau-Aduli, B. Systematic review of specialist selection methods with implications for diversity in the medical workforce. 2021. *BMC Medical Education*. 21:448.

This chapter describes the methodology and findings of a systematic review of the literature on methods used to select candidates for entry into training programmes for specialist areas of medicine, such as psychiatry, surgery, or general practice (each of which may be further divided into subspecialties, such as adult psychiatry or neurosurgery). Every specialist medical training programme requires the development of a specific curriculum, and the selection of candidates is guided by that curriculum. This chapter evaluates the evidence that current methods of specialist training selection include biases that explain the relative under-representation of groups including

women and ethnic minorities. The production of empirically derived maps of specialist medical knowledge would be one way of addressing such biases.

3.1 - Abstract

Purpose

There is growing concern that inequities in methods of selection into medical specialties reduce specialist cohort diversity, particularly where measures designed for another purpose are adapted for specialist selection, prioritising reliability over validity. This review examined how empirical measures affect the diversity of specialist selection. The goals were to summarise the groups for which evidence is available, evaluate evidence that measures prioritising reliability over validity contribute to under-representation, and identify novel measures or processes that address under-representation, in order to make recommendations on selection into medical specialties and research required to support diversity.

Method

In 2020-1, the authors implemented a comprehensive search strategy across 4 electronic databases (Medline, PsychINFO, Scopus, ERIC) covering years 2000-2020, supplemented with hand-search of key journals and reference lists from identified studies. Articles were screened using explicit inclusion and exclusion criteria designed to focus on empirical measures used in medical specialty selection decisions.

Results

35 articles were included from 1344 retrieved from databases and hand-searches. In order of prevalence these papers addressed the under-representation of women (21/35), international medical graduates (10/35), and race/ethnicity (9/35). Apart from well-powered studies of selection into general practice training in the UK, the literature was exploratory, retrospective, and relied upon convenience samples with limited follow-up. There was preliminary evidence that bias in the measures used for selection into training might contribute to under-representation of some groups.

Conclusions

The review did not find convincing evidence that measures prioritising reliability drive under-representation of some groups in medical specialties, although this may be due to limited power analyses. In addition, the review did not identify novel specialist selection methods likely to improve diversity. Nevertheless, significant and divergent efforts are being made to promote the evolution of selection processes that draw on all the diverse qualities required for specialist practice serving diverse populations. More rigorous prospective research across different national frameworks will be needed to clarify whether eliminating or reducing the weighting of reliable pre-selection academic results in selection decisions will increase or decrease diversity, and whether drawing on a broader range of assessments can achieve both reliable and socially desirable outcomes.

Keywords:

Diversity; Justice; Equity; Specialist selection; Residency; Bias; Gender; Ethnicity; Application; Matching

3.2 - Background

There is long-standing recognition that medical workforces do not represent the diversity of the populations they serve.[1] While there have been improvements in the representation of some under-represented groups, particularly women, as a proportion of medical students and junior doctors, significant imbalances remain among senior doctors and competitive specialties.[1-5]

The pattern of under-representation of racial and ethnic minorities is more variable than gender, but equally concerning. One report noted that African Americans, Hispanic Americans, and American Indians comprised more than a quarter of the US population but only 6% of its physicians.[1] The same report argued that increased diversity of the health workforce was justified both to support social justice, and as an effective means of improving population health by improving cultural competence, communication, patient trust, and reducing barriers to care.[1,6] In response to similar concerns, some

medical schools have developed socially accountable education frameworks where community collaboration, equitable selection criteria not solely focused on academic performance, and learning experiences in areas of need are used to encourage recruitment and retention to rural and other underserved populations.[7]

Despite the importance of racial and ethnic diversity in the medical workforce there has been less progress in these groups than gender.[5,8,9] The barriers to medical workforce diversity are varied, but can be summarised as due to differential resources, selection bias, and anticipated bias,[10] leading some to conclude that bias may be reduced if examiners have similar demographics to candidates.[11]

A variety of historical and current conditions mean that under-represented minorities (URMs) have fewer material and cultural resources than privileged groups to match the challenges associated with preparing for application to medical school, and for navigating the pathways through medical training to specialist practice.[1] Although it has been argued for some time that the focus on academic performance ignores many of the qualities which contribute to competent, caring, and ethical medical practice,[12] there has been little progress in developing and implementing reliable non-academic indicators of aptitude for medical practice.[13] As Roberts et al[14] make clear, all current methods of selection into medical specialty training may contribute to biased selection. The most reliable instruments used for selection into medical specialties are multiple choice question (MCQ) tests, because the format allows for a large number of items and a broad coverage of content. Efforts to improve the validity of selection decisions are less well developed, although there has been an effort in the UK to improve the validity of selection decisions by developing a suite of reliable measures across a range of relevant skills and knowledge.

Biased measures during trainee selection may be one cause of under-representation of some groups in medical specialties, tending to favour privileged groups.[14] For example, men have shown a small but reliable advantage over women on the MCQ tests used for medical school selection, while women have shown an advantage on the clinical assessments performed during medical school.[15]

Perhaps anticipating this type of selection bias, or as a result of differential resources, URMs may be less likely to apply for medical school or specialist training than other people with similar levels of ability.[16]

3.2.1 - The broader medical training selection literature

Useful context is provided by two recent reviews which describe a tension between the reliability and validity of the processes and instruments used for selection along the training trajectory from medical school through to consultant practice. After canvassing the significantly different trajectories in different countries through medical school, selection into generalist training, and transition to consultant practice, Roberts et al.[14] propose two basic national patterns of medical specialty training selection (MSTS) with the US representative of a pattern of relatively greater dependence upon pre-selection academic achievement combined at the local level with subjective measures such as letters of recommendation; and the UK in the early stages of developing a systematic framework that combines multiple reliable methods of selection covering a broad range of skills.

The heavy reliance of the US MSTS framework on pre-selection academic achievement is illustrated by the status of the United States Medical Licensing Exam - Part I (USMLE I) as the most common tool used for MSTS in the US, despite being created for licensure as a doctor at the end of medical school.[14] The USMLE I is very attractive to administrators responsible for MSTS decisions because of its convenience as a reliable, standardised, pre-existing measure allowing the direct comparison of a large majority of US doctors on a measure of characteristics ostensibly relevant to specialist practice without the need for additional testing. These benefits are so significant that they overwhelm the questionable validity of using the same test to select into specialties as diverse as psychiatry, surgery, and paediatrics, and in fact have been argued to have prevented the development of more valid measures targeting specific specialties.[17]

This tension between reliability and validity, with the strong temptation to focus on reliability for its administrative convenience, is an example of the long-recognised problem that focusing management only on what is most conveniently measured ignores crucial factors which may not be so easily measured.[13,18] Social accountability theory suggests that selecting candidates for entry into medical school or medical specialties based purely on pre-selection academic achievement is likely to ignore many socially important goals, often exacerbating existing inequities.[19]

Due to the overlapping methods and analysis, and the larger dataset, further context is available from Patterson et al's[20] review of the methods of selection into medical school. They conclude that the validity and reliability of selection decisions may be improved by developing specific measures using structured techniques such as situational judgement tests (SJTs) and multiple-mini interviews (MMIs) (both described in Table 3-1), while the greater reliability of pre-selection academic achievement measures may involve the cost of preventing the entry of some under-represented minorities into medical training. Both these reviews illustrate the over-reliance of medical selection research on retrospective, cross-sectional designs and the tendency to focus on reliable more than valid indicators. While a full exploration is beyond the scope of this review it is useful to note that the tension between reliability and validity is important outside the boundaries of academic medicine. The large size and crucial social functions played by health workforces makes their composition a live political issue, leading to calls for the reduction of the reliance on standardised tests to improve the diversity of selection into health professions more generally, which may be interpreted as a restatement of the tension between reliability and validity translated into more commonly understood language.[1,21]

Table 3-1 – Common instruments for selection into medical specialist training programmes[14,20]

Instrument	Description
Interviews/Multiple mini-interviews	Includes standardised and non-standardised interviews, which may be supported by psychometric evidence, although frequently involve subjective judgements.
Academic records	Particularly school results measured against a year-cohort, but may include other information, such as extra-curricular activities, awards, etc
Standardised exams/aptitude tests (including SJT/CPST)	Includes exams which test general medical, not specialist, aptitude: • Standardised exams used for selection into medical school or licensure for practice, such as the United States Medical Licensing Exam(s) and the UK's Multi-Specialty Recruitment Assessment And exams designed for particular specialties, including: • OSCE format interviews • Situational judgement tests which assess non-cognitive characteristics by presenting workplace-based scenarios requiring non-clinical decisions • Clinical problem-solving tests (CPST) which involve multiple-choice responses to clinical scenarios requiring clinical reasoning
Curriculum vitae	Structured or free-form document(s) provided by candidate outlining their education, training, and work experiences.
Letters of recommendation	Structured or free-form letters expressing an opinion on the candidates' specific or general capacities, often weighted for the perceived expertise

	or prestige of the undersigned; for example greater weight may be given to a LoR by the Dean of a prominent medical school than a consultant in a medical specialty.
Personal statements	Structured or free-form statements by the candidate usually addressing specific criteria such as motivation, priorities, and personal circumstances.
Referees reports/references	Structured or free-form reports by referees with knowledge of the candidate addressing specific selection criteria.
Locally defined criteria	The criteria used for selection into individual specialist training programmes may not be precisely defined. Locally defined criteria may involve algorithms weighting various of the instruments described above, and may or may not involve objective thresholds or subjective judgements

3.2.2 - Review goals

In the context of the tension between the reliability and validity of MSTS measures and the pragmatic advantages of reliable measures, this article was designed to review and evaluate the research on how MSTS instruments affect the diversity of selection into medical specialty training programs, and make recommendations for balancing the goals of reliable and equitable MSTS, justifying the following research questions:

- What URMs have been considered regarding the impact of empirical MSTS methods on diversity?
- What research designs have been used to examine the impact of empirical MSTS methods on diversity?

- What evidence suggests that reliance on measures of pre-selection academic achievement decrease MSTS diversity?
- What evidence suggests that novel selection processes improve diversity relative to pre-selection academic achievement measures and what is their impact on reliability?

3.3 - Method

3.3.1 - Study selection

Study inclusion/exclusion criteria are presented in Table 3-2. To focus on the effect of specific measures used in the decision to accept candidates into specialty training, studies which reported surveys or other ways of measuring candidate perceptions, motivations, and preferences were excluded. Table 3-1 describes the common instruments used for selection in the literature.

Table 3-2 – Inclusion and exclusion study criteria

Inclusion criteria	Exclusion criteria

• Selection into medical specialty training program
• Results report empirical evidence about a measure used for medical specialty selection
• Focus of article is on diversity or under-represented minority in medical specialty training
• Published between 1.01.2000 and 31.12.2020
• English
	• Selection into medical school
	• Selection into non-medical training:
	  ○ Nursing
	  ○ Allied Health
	  ○ Dental
	  ○ Pharmacy
	• Articles where diversity or underrepresented minority in medical specialty training is not the focus
	• Not in English
	• Published prior to 1.1.2000 or after 31.12.2020
	• Survey results only

	• Empirical results relate only to preferences, perceptions, motivations to apply, and not measures used as basis of selection

3.3.2 - Search strategy

The search was based on the method suggested by Aveyard.[32] Searches were repeated in PubMed/Medline, PsycINFO, Scopus, and ERIC, in order to identify relevant articles from the medical, psychological, and educational literature (see search strings in Supplementary materials). Search results were supplemented with hand-search of key journals, articles in the reference lists of the articles selected for inclusion in the review, and articles which cited the articles selected for inclusion in the review (identified using Web of Science). Key journals were defined as those with two or more articles selected for review, including: Medical Education, BMC Medical Education, and Academic Medicine.

During the search, the terms used for doctors in medical specialty training included “resident”, “trainee”, and “postgraduate”. Where specific instrument or minority search terms were added to the basic search, they were added as “OR” clauses that would return a larger set, and never used to constrain/reduce searches. Such additional search terms referred to specific instruments of selection used in the US (United States Medical Licensing Exam – USMLE; of several parts USMLE 1 and USMLE 2 are commonly used for selection) and the UK (SJT – Situational Judgement Test, CPST – Clinical Problem Solving Test). The two most common URMs, gender and international medical graduates, were also specifically added. A broad net was cast for articles about diversity including the terms divers*, equit*, gender, foreign, international, underrepresented, and minority.

3.3.3 - Data extraction and analysis

Each article was reviewed with reference to a standard data extraction pro-forma designed for this study (see Supplementary materials). An excel spreadsheet collected and summarised information from the pro-forma. Methodological strengths and limitations were systematically collected and coded in relation to scope of study, research quality, sample size, power analysis, specialty and length of study/ follow-up.

We used the Medical Education Research Study Quality Instrument (MERSQI) as a standardised measure of article quality.[33,34] This instrument covers six domains comprising study design, sampling, type of data, validity of evaluation instrument, data analysis, and outcomes measured, with scores varying between 5 and 18. Two of us (AA & BMA) independently completed the MERSQI for each article, and resolved disagreements with reference to MERSQI criteria in a joint session, achieving consensus. A recent review of studies using the MERSQI to assess the quality of medical education studies reported a range of overall scores between 8.9 – 15.1 (max 18) with a median of 11.3, while recommending that quality should also be assessed by examination of the specific features and conditions of individual studies.

3.3.4 - Post-hoc analysis of unbalanced results

In response to the search results, with a single article (from Canada) outside the dominant set from the US and a smaller set from the UK, it was decided to analyse what impact the use of specific search terms including instruments used primarily in the US (USMLE) and UK (SJT/CPST) and specific minority groups (gender/IMG) had on the search results. As we used specific terms only to increase the number of hits and not to decrease them, we do not think it was possible to have introduced a bias against finding research with particular characteristics (such as research done outside the US/UK). However, it seems possible that using specific search terms could have misrepresented the literature by tending to return a greater proportion of US/UK and/or gender/IMG articles. We tested this in two ways: to examine whether we might have missed additional articles eg from other

countries, we extended our search over the 2000-2020 time period, to a fourth database, Scopus, the largest database available to us. To quantify the potential bias of having a greater probability of identifying articles from US/UK than elsewhere we identified the articles which were included in our review which were not identified by our basic search, but which were added as a result of the specific search terms above.

3.4 - Results

The database searches retrieved a total of 1344 abstracts with 1275 unique articles after 69 duplicates were removed (Figure 3-1). 18 articles were added after the hand-search of key journals and reference/citation review. Application of the inclusion/exclusion criteria identified 64 articles for full-text retrieval, and full-text review yielded 35 articles for inclusion in the article.

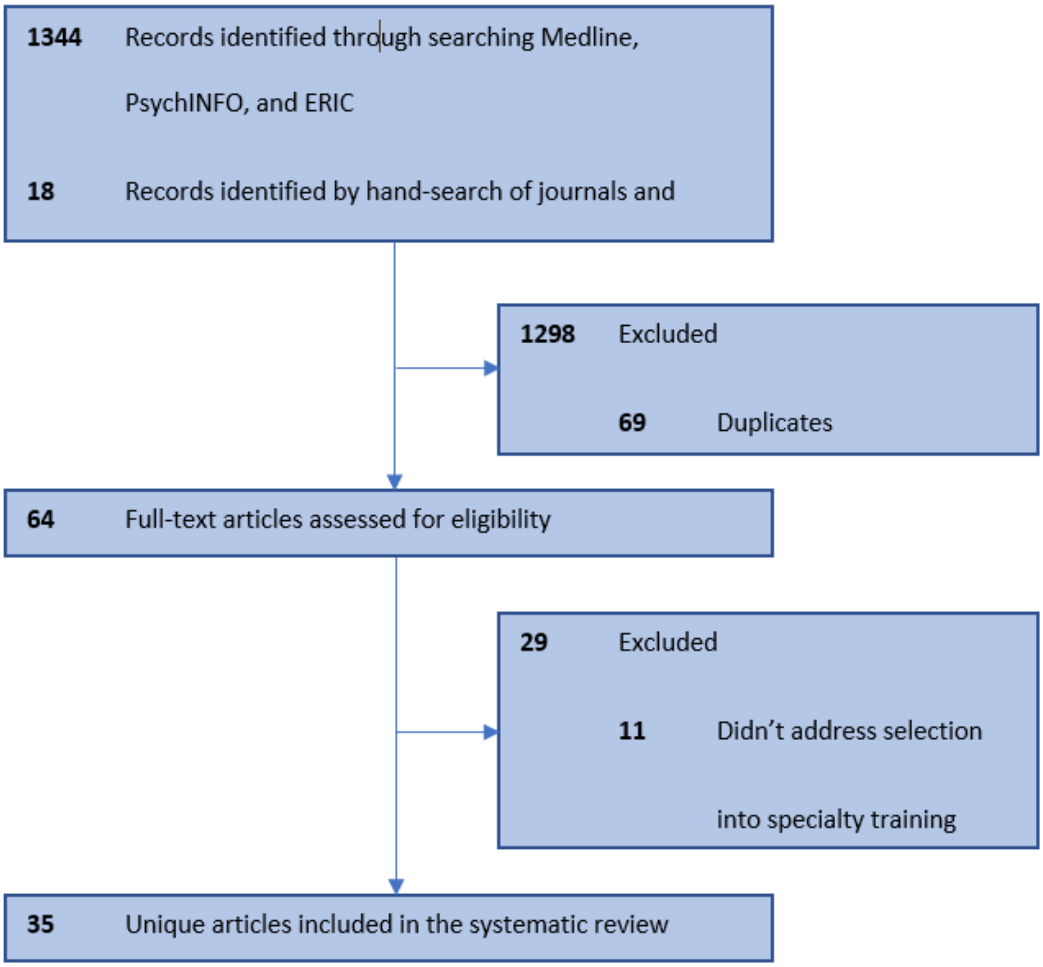


Figure 3-1: PRISMA Flowchart of literature search and article inclusion/exclusion

The retrieved articles comprise a heterogeneous set with few commonalities, described below and summarised in Table 3-3.

Table 3-3 – Summary of Reviewed Articles

Article	Description	Main findings	Diversity conclusions (shading indicates claimed bias)	Strengths/limitations	MERSQI Score[a] (11.3/18 over all articles)
Canada (1 article)					9
MacLellan et al (2010)	Compared IMG and DMG performance on in- and end-training exams	End-training exam pass rate: IMG 56% versus DMG 93.5% (p<.0001)	**IMG:** IMG low pre-selection scores consistent with low pass rates on certification exams	**Strengths:** Multiple year, large sample **Limitations:** Exploratory, single program, single specialty	9
UK (7 articles)					15.2
Esmail et al (2013)	Compared IMG with DMG performance on	URM failed first attempt more than white DMG (OR 3.5, p<.001)	**URM/IMG:** Higher failure rates in domestic and foreign URM/IMG	**Strengths:** Complete cohort, large sample,	15.8

	end-training exams (GP/Family medicine)	IMG failed first attempt more than white DMG (OR 14.7, p<.001)	are partly explained by lower pre-selection academic achievement, and may also reflect bias during clinical OSCE-based exams	multiple years, end-training outcome **Limitations:** Exploratory, single specialty	
McManus et al (2014)	Compared IMG with DMG performance on end-training exams (GP/Family medicine & Internal medicine)	IMG performed worse than DMG on end-training exams (~1.25 SD)	**IMG:** Lower pre-selection scores are an accurate measure of suitability for training Raising cutoffs is needed for equivalence with DMG but would affect workforce	**Strengths:** Follow-up study, multiple programs, large sample, multiple years **Limitations:** Two specialties	15.2

Patterson et al (2018)	Measured factors associated with differences in performance of IMG and DMG on end-training exams (GP/Family medicine)	Clinical skill performance better predicted by SJT than CPST (beta 0.26 v 0.17) SJT mediated relationship between English fluency and clinical skills performance	**IMG:** IMG performance on end-of-training exams is predicted by socio-linguistic factors not clinical knowledge and skills	**Strengths:** National cohort, large sample, multiple years, end-of-training follow-up **Limitations:** Exploratory study, single specialty	14.6
Tiffin et al (2014)	Measure IMG performance during residency	IMG more likely to receive unsatisfactory ARCP than DMG (OR 1.63, p<.05)	**IMG:** PLAB language exam does not establish linguistic equivalence of IMG and DMG Thresholds would need to be increased to achieve equivalence, but would affect	**Strengths:** National cohort, large sample Limitations:	14.6

			workforce and decrease diversity		
Tiffin et al (2018)	Measure bias against IMG in resident selection comparing pre-training academic attainment with in-training assessment	UK overseas graduates more likely deemed appointable than IMG (OR 1.29, p<.05) but more likely to later receive less satisfactory ARCP (OR 1.20, p<.05)	**IMG:** Bias favouring UK born graduates trained overseas versus IMGs may be due to excessive weight given to interview	**Strengths:** National cohort, large sample, all specialties, **Limitations:** Incomplete data set	15.8
Wakeford et al (2015)	Measure correlation between GP/Family medicine and Internal medicine exam performance by ethnicity	High correlation between GP/IM exam performance, suggesting validity of each assessment (and does not suggest bias against URM) URM performed less well	**URM:** No evidence of bias against URM; differences in assessment likely to reflect true differences in ability	**Strengths:** National cohort, multiple years, large sample **Limitations:** Exploratory, two specialties	15.8

Woolf et al (2019) Identified by specific search terms	Measure effect of gender on specialty training selection	Across all specialties female applicants had: • No difference in applications • Increased offers (OR 1.4, p<.001) • Increased acceptance (OR 1.43, p<.001) 2 specialties had significant gender differences in applications (both favouring women): • Paediatrics (OR 1.57, p<.05)	**Gender:** Gender segregation in specialties is due to differential application rates, not instrument bias; research is needed on why men are less likely to apply for GP/Paediatric training, and less likely to accept GP training if offered	**Strengths:** Follow-up study, national cohort, large sample, multiple specialties **Limitations:** 1-2 years intake, incomplete data set	14.6

		• GP (OR 1.23, p<.05)			
US (27 articles)					10.4
Aisen et al (2018) Identified by specific search terms	Examine effect of gender on **urology** applicant academic achievement and selection into specialty	Higher % of males matched (73% v 67%) Among matched applicants: • Males less honors (2.8 v 2.2, p<.021) • Males higher USMLE1 (245.9 v 240.8, p<.001)	**Gender:** Male/Female candidates had similar pre-selection results and no evidence of bias in selection	**Strengths:** Moderate size **Limitations:** Exploratory, single program, single specialty, 1-2 years intake	11.3
Brandt et al (2013)	Examine effect of gender on **O&G** applicant academic	No gender difference on USMLE Females more likely to have honors (51% v 41%, p<.021)	**Gender:** Male/Female candidates had similar USMLE1 scores, higher female honors may explain lower rate of M	**Strengths:** Large sample, multiple years **Limitations:** Exploratory, single program, single	11.3

	achievements and selection into specialty	and published (87% v 79%, p<.01)	applications for O&G training	specialty, incomplete data set	
Chapman et al (2019)	Identify factors associated with under-representation of women across **medical specialties**	Female representation higher in specialties with lower mean USMLE1 entry score (p<.017) 1% increase in female faculty prevalence associated with 1.45% increase in female trainees in specialty (p<.001)	**Gender:** No evidence of USMLE 1 bias against females Association between female faculty and female trainees suggests mentoring may increase diversity	**Strengths:** National cohort, large sample, all specialties **Limitations:** Exploratory, 1-2 years intake, incomplete data set	9
De Oliveira et al (2012) Identified by specific search terms	Measure factors associated with selection to **anaesthetics** residency	Factors associated with selection: • Female • Younger • Higher USMLE 2	**Gender/Age:** Bias favouring selection of **female** and **younger** applicants	**Strengths:** Large sample **Limitations:** Exploratory, single program, single specialty, 1-2 years intake,	12.4

	including gender, age, country of training	• DMG		inferences made without statistical test	
Dirschl et al (2006) Identified by specific search terms	Measure whether gender and academic scores can predict **orthopaedic** end-of-training exams	12.5% female applicants Faculty ratings of training were not associated with academic scores	**Gender:** No gender bias detected	**Strengths:** Follow-up study, large sample, multiple years **Limitations:** Single program, single specialty	9
Driver et al (2014)	Identify factors associated with **ophthalmology** selection including IMG status	Increased % of selection associated with: • Higher USMLE1 (OR 3.22, p<.05) • Letters of recommendation (OR 6.2, p<.05)	**IMG:** Design prevented conclusions about bias	**Strengths:** National cohort, large sample, multiple years **Limitations:** Exploratory, single specialty	11.3

		• Publications (OR 3, p<.05)			
Durham et al (2018)	Measure effect of gender on selection into **neurosurgical** training	13.8% female applicants USMLE1 higher for selected (233 v 211, p<.001) Females had lower OR of matching (0.59, p<.001) Females had lower mean USMLE1 scores (222 v 230, p<.001)	**Gender:** USMLE 1 is best predictor of selection **Reduced female** selection partially explained by lower USMLE 1 scores Possible bias remains after multivariate analysis	**Strengths:** Statewide cohort, large sample, multiple years **Limitations:** Exploratory, single specialty	11.3
Edmond et al (2001)	Measure bias against African Americans due to **USMLE 1** in **internal**	Mean USMLE1 of African Americans was 200, non-AA was 216	**Race:** USMLE 1 reduces selection of **African Americans**	**Strengths:** Large sample **Limitations:** Exploratory, single program, single	12.4

Identified by specific search terms	**medicine** residency selection	OR for rejection of AA varied from 3 – 6 (p<.05)		specialty, 1-2 years intake, uncontrolled confound	
Filippou et al (2019)	Measure gender bias in letters of recommendation for **urology** resident applicants	LoR for males had: • More authentic tone • More references to personal drive, work, and power LoR referring to power more likely to be associated with selection	**Gender:** Gender bias in letters of recommendation may reduce selection of **females**	**Strengths:** Moderate sample **Limitations:** Exploratory, single program, single specialty, 1-2 years intake	9
French et al (2019)	Measure gender bias in LoR for **general surgery** resident applicants	Female authors wrote longer letters	**Gender:** No gender bias detected in letters of recommendation	**Strengths:** Large sample, adequate power	7.9

				Limitations: Exploratory, single program, single specialty, 1-2 years intake	
Friedman et al (2017)	Measure gender bias in standardised versus narrative LoR for **otolaryngology surgery** residents	No difference in ranking of male/female applicants Female writers produce LoRs different to male writers ($p<.05$) LoRs written for female applicants less positive than those written for male applicants ($p<.05$)	**Gender:** Standardised letters of recommendation have reduced but not eliminated biases that contribute to reduced selection of **females**	**Strengths:** Moderate sample **Limitations:** Exploratory, single program, single specialty, 1-2 years intake	7.9
Gardner et al (2019)	Measure effect of USMLE cutoffs on underrepresented	Reducing USMLE1 cutoffs and adding SJT screening	**Gender/URM:** USMLE 1 screening reduces	**Strengths:** Multiple program sample, large sample	9

	minorities in **general surgery** training	increased URMs offered interview by 8%	selection of URMs for interview Does not claim bias	**Limitations:** Exploratory, single specialty, 1-2 years intake	
Girzadas et al (2004)	Measure effect of gender on SLoR for **emergency medicine** residency	Female author with female applicant OR 2 to get highest ranking on LoR (p=.023)	**Gender:** No gender bias detected in letters of recommendation	**Strengths:** Large sample **Limitations:** Exploratory, single program, single specialty, 1-2 years intake, selection process changed during study	7.9
Hewett et al (2016)	Measure gender bias in **radiology** residency selection	24% female applicants Females were • 30% of offered interviews	Gender: Bias favouring female applicants Associated with lower female USMLE1 scores	**Strengths:** Multiple years intake, large sample **Limitations:** Exploratory, single program, single	11.3

		• 38% of top quartile (p<.001) • 25% of selected Female applicants average USMLE1 score was 5 points lower (p<.05) Female applicants had higher mean interview scores (p<.05)	Associated with higher female interview scores	specialty, variable selection/scoring methods	
Hoffman et al (2020)	Measure gender bias in LoR for **pediatric surgery** residency selection	Female LoR had more communal phrases (p<.01)	**Gender:** Gender biases **against females** in LoRs may affect selection into training	**Strengths:** Multiple years intake **Limitations:** Exploratory, single program, single specialty, small sample, ad-hoc measures	7.9

Hoffman et al (2019)	Measure gender bias in LoR for **transplant surgery** resident applicants	Male applicant LoR had more agentic terms (p<.05) LoR written by senior staff more likely to describe female applicants with communal terms (p<.05)	**Gender:** Gender biases in LoRs **against females** may affect selection into training	**Strengths:** Moderate sample size, multiple years intake **Limitations:** Exploratory study, single program, single specialty, limited power	7.9
Hopson et al (2019) Identified by specific search terms	Measure influence of gender on outcome of **emergency medicine** selection interviews	No significant difference on standardised video interview	**Gender:** No gender bias detected on standardised video interview	**Strengths:** Multiple program cohort, large sample size, adequate power reported **Limitations:** Exploratory study, single specialty, 1-2 years intake, aggregates	10.1

				heterogenous groups, ad-hoc measures	
Kobayashi et al (2019)	Measure influence of gender on LoR in **orthopaedic** surgery residency	Female applicants had: • Longer LoR (p<.003) • More "achieve" words (p<.0001) No differences for male v female authors	**Gender:** No gender bias detected on letters of recommendation	**Strengths:** Large sample **Limitations:** Exploratory study, single program, single specialty, 1-2 years intake, ad-hoc measures	11.3
Lin et al (2019)	Measure gender bias in LoR for **ophthalmology** residency	M/F applicants had similar: • USMLE1 • Academic achievement LoR for male applicants had:	**Gender:** Gender biases in LoRs **against females** may affect selection into training	**Strengths:** Moderate sample size **Limitations:** Exploratory, single program, single specialty, 1-2 years intake, ad-hoc measures	11.3

		• Less feel words (p<041) • Less biological words (p<.028)			
Lypson et al (2010) Identified by specific search terms	Measure correlation between USMLE scores and clinical competence at beginning of residency **across specialties**	USMLE1 scores lower for URM (212 v 230, p<.001) URM not significantly worse than non-URM on OSCE stations at beginning of residency	**URM:** USMLE 1 scores are biased against URMs, revealed by similar OSCE scores at beginning of residency	**Strengths:** Multiple specialties, multiple years intake **Limitations:** Exploratory, single program, small sample, limited power	7.9
Norcini et al (2014)	Predict patient outcomes of IMGs from USMLE scores **across specialties**	Increased USMLE2 CK score associated with decreased mortality as a physician	**IMG:** USMLE2 CK scores are a valid measure of suitability for IMG selection/certification	**Strengths:** Follow-up study, statewide sample, large sample, multiple specialties, multiple years intake, patient outcomes	14.5

		1 SD on USMLE 2 CK associated with 4% improvement in mortality		**Limitations:** Unmeasured confounds	
Poon et al (2019) Identified by specific search terms	Compare **orthopaedic** residency enrolment rates and academic metrics of applicants and matriculated residents by race/ethnicity	URM were 29% of applicants and 25% of enrolments White/Asian applicants had higher USMLE1 than Black applicants (234 v 218, p<.05)	**URM:** USMLE1 screening may contribute to lower rates of application of URMs Bias not evaluated	**Strengths:** National cohort, large sample, adequate power **Limitations:** Important variables not measured	13.5
Quintero et al (2009)	Measure effect of personality similarity to bias the selection of **orthopaedic** residents	Clinicians rated candidates more favourably when they shared personality characteristics (p=.044)	**Personality:** Increased awareness of implicit biases may reduce inequity of current selection processes	**Strengths:** Moderate sample size **Limitations:** Exploratory, single program, single specialty, 1-2 years intake,	12.4

				limited power, follow-up to selection, protocol variations	
Scherl et al (2001)	Measure gender bias in **orthopaedic** resident selection	No significant difference in selection of male and female charts	**Gender:** No gender bias detected based on gendered versions of applicant charts	**Strengths:** Experimental design **Limitations:** Exploratory, single program, small sample, selection bias, partial blinding	11.3
Stain et al (2013) Identified by specific search terms	Measure attributes of top-ranked applicants to **general surgery** residency	Males had higher USMLE1 (238 v 230, p<.001) Males/Females had similar USMLE2 scores (245 v 244, p=.54)	**Gender:** No gender bias detected based on pre-selection academic achievements	**Strengths:** National cohort, moderate sample size **Limitations:** Single program, single specialty, ad-hoc measures	12.4

		Highly competitive programs associated with • USMLE1 (RR 1.36) • Publications (RR 2.2) • Asian (RR 1.7 v white)			
Unkart et al (2016)	Measure reduction in **general surgical** residency applications among candidates self-identified as "disadvantaged"	URM were: • Older at entry (24 v 23, p<.001) • Lower MCAT (30 v 33, p<.001) • More likely to choose a less competitive specialty (p<.03)	**URM/Gender:** No bias detected based on USMLE 1	**Strengths:** National cohort, multiple years intake, large sample **Limitations:** Aggregates heterogenous groups, limited follow-up	12.4

Villwock et al (2019) Identified by specific search terms	Measure effect of STAR tool for selecting **otolaryngology** residency candidates to interview	USMLE scores significantly increased after STAR tool No differences in gender/URM before/after introduction of STAR selection tool	**URM/Gender:** STAR selection tool did not increase representation of URM/Gender	**Strengths:** Moderate sample size **Limitations:** Single program, exploratory	7.9

a – MERSQI scores include subscales which are not applicable for all articles; scores are scaled after removal of these subscales to allow comparison with a maximum score of 18 for all articles (Reed et al, 2007)[17]

ARCP – Annual Review of Competence Progression, CPST – Clinical Problem Solving Test, DMG – Domestic Medical Graduate, IMG – International Medical Graduate, LoR – Letter of Recommendation, PLAB – Professional and Linguistic Assessment Board, SJT – Situational Judgement Test, URM – Underrepresented minority

3.4.1 - Under-represented minorities

Gender was by far the most frequently examined URM (22/35 articles: 62%), followed by international medical graduates (IMGs) (10/35: 28%). Nine articles reported multiple classes of URM (26%) and single articles considered age,[35] personality,[36] and geography[37] (each 3%).

3.4.2 - Methods used to investigate diversity of selection

Most of the studies were conducted in the US (27/35 articles; 77%) and after 2013 (24/35; 69%), with smaller contributions from the UK (7/35; 20%) and Canada (1/35; 3%). Surgery (18/35; 51%) and GP (5/35; 14%) generated the most articles of any single specialty, with most of the other specialties contributing one or no specific articles.

Table 3-3 summarises the strengths, limitations, and MERSQI scores of each article. The mean MERSQI score was 11.34 (SD: 2.61; range: 7.9-15.8) which is comparable with the previous literature using MERSQI as a measure of study quality. Across all articles, mean MERSQI scores were adequate for all domains except study design (1.25 out of 3) and data analysis (1.5 out of 3). The interrater reliability across all domains was in the fair (0.21 – 0.4) or moderate (0.41 – 0.6) range (Cohen's Kappa) except where a lack of variation in the coded scores prevented calculation.

Consistent with the MERSQI scores of previous studies, closer examination of the collected articles revealed significant methodological limitations particularly in design and analysis (Table 3-3). Critically, a substantial minority only considered applicants that had already been selected into a training program, not those who were unsuccessful (26%). Prevalent limitations of the literature include that most of the articles were exploratory in nature (83%), and examined a single training program (56%), or a single specialty (78%).

Many articles had the strength of looking at a complete training cohort across a nation or state (34%), and most of the studies used large sample sizes (>500 candidates; 69%). Across 35 articles, data was reported on 200,000 participants, with the UK articles averaging more than 17,000 participants and the US more than 2700 per article. Most of the studies also examined selection over multiple intake cycles (54% of articles considered more than 2 years of data). In contrast with the exploratory US literature, the 7 UK articles were part of a coordinated research effort using similar methods on national data sources focused on GP training and with a greater interest in the reliability of assessment of IMGs than other URMs.

While the methods, populations, and quality of the studies were too heterogeneous to allow meta-analysis, power was examined as a useful index of the quality of the research. Reflecting the primarily exploratory nature of the research, 17% of articles reported adequate power, 8% reported limited power, and 74% did not address power.

Also consistent with the exploratory nature of the research, most of the articles relied on retrospective cohort studies (89%), with only three prospective studies. Pre-selection academic achievement comprising MCQ exams were considered by most of the articles (74%), followed by letters of recommendation (33%), and a small number examining standardised or non-standardised interviews (8%) and selection centres (8%; sum greater than 100% as some studies looked at more than one selection method). Figure 3-2 shows that most of the literature had a limited follow-up period, with most articles considering only the process of application to training (15/35) or selection into a specialty (10/35). Few articles considered the impact of selection processes on in-training assessment (4/35) or certification exams (5/35), and only one looked at the effects of selection on consultant practice.

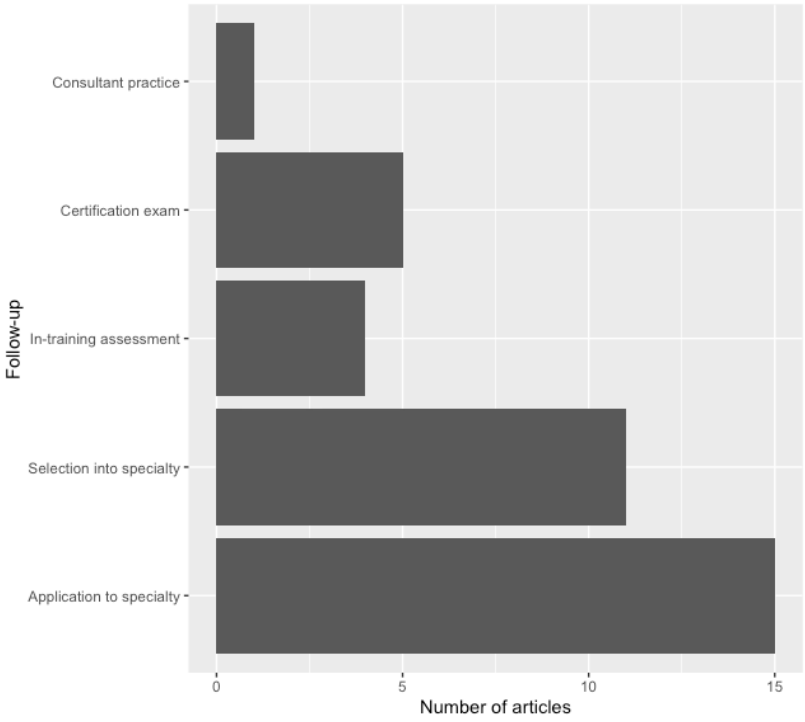


Figure 3-2: Length of follow-up

3.4.3 - Impact of pre-selection measures on diversity

Table 3-3 summarizes the impact of pre-selection measures on MSTS (shading indicates evidence of bias likely to affect diversity). The lone Canadian article found no evidence of bias against IMGs. Three of seven UK articles concluded there was evidence of bias against URMs or IMGs. Eleven of twenty-seven US articles found evidence of bias, with two showing greater selection of women due to better performance on the USMLE 2 and interview; one showing lesser selection of women associated with worse performance on the USMLE 1; and five showing bias against women on letters of recommendation. None of the other significant results were supported across more than one study.

3.4.4 - Evidence that novel selection processes can increase diversity of selection

Two articles reported evidence on novel selection processes designed to increase diversity of selection. Gardner et al[38] found that reducing the threshold of the USMLE 1 and adding a SJT with MCQs specifically designed for selection into surgical training increased the selection of URMs for interview by 8%. Villwock et al[37] reported that an objective algorithm for selecting candidates for interview (Selection Tool for Applicants to Residency – STAR), designed to prevent unconscious bias with attention to multiple factors including geographical (eg candidates proximity to the selecting institution), did not increase the proportion of URMs offered interviews for otolaryngology training.

3.4.5 - Potential bias attributable to search strategy

Our replication of the basic search in the Scopus database did not identify any additional articles for review. Table 3-3 indicates which of the reviewed articles were identified by the addition of specific search terms to our basic search. 10 articles of the 35 reviewed were not retrieved by the basic search, of which 3 reported evidence of bias. The US literature provided 9 of the additional 10, with the other from the UK.

3.5 - Discussion

3.5.1 - Summary of findings and similarity to previous literature

The MSTS diversity literature focused mainly on under-selection of females into specialist training, followed by IMGs and then race or ethnicity. Apart from a small group of high quality studies from the UK with adequately powered large samples from national cohorts as part of the development of a systematic framework for GP trainee selection, evidence was limited by exploratory retrospective designs using convenience samples of single specialties and single training programs, with brief follow-up periods. Alongside the methodological limitations of the individual studies in this review,

the large variations in the frameworks for MSTS between specialties within the same country, and even greater variations across countries, makes it difficult to draw confident conclusions from this literature. The results are consistent with recent reviews of medical school and specialty selection methods[14,20] both in the dominance of US research with a smaller but more coherent set of articles from the UK; and with respect to their conclusions that reforming selection frameworks to achieve reliable and equitable selection will require research with greater methodological rigour, particularly longitudinal design and attention to validity.

Perhaps reflecting the relatively low diversity in surgical programs,[14] half the studies examined one of the surgical subspecialties. Outside the GP focus of the UK literature, most non-surgical specialties were represented by a single article, or not represented at all. There was equivocal evidence of bias against the selection of females into specialist training, and contested evidence of bias against IMGs. The use of specific search terms in addition to the baseline search did not exclude any articles from review, but did identify an additional 10 articles, primarily from the US literature. The additional evidence reviewed appears unlikely to have significantly altered the analysis, conclusions, or recommendations of the review. Given the similar results of a recent review of MSTS not focused on diversity we believe our review is representative of the published literature.

3.5.2 - Methods used to investigate diversity in medical specialty selection

Although the methods used and populations sampled were diverse, almost all articles had retrospective cohort designs, and most of the research only followed up to the point of selection into training, with few looking as far as in-training assessments or certification exams. Durham et al[39] is representative. They found that the USMLE 1 was the best predictor of selection into US neurosurgical training across all candidates. While reduced female selection was partially explained by lower USMLE 1 scores, multivariate analysis suggested that women were less likely to be selected even after controlling for the USMLE and other academic measures, which was interpreted as

evidence of possible gender bias. This study shows two potential barriers to selection of female trainees: lower average USMLE 1 scores, which the authors implicitly accept as a reasonable index of ability; and gender bias of the whole selection process, which they do not consider acceptable.

It is notable that 26% of articles only reported data on people already selected into training. While these studies can compare URMs and others selected into training, it is difficult to explain barriers to MSTS without data about URMs who have been excluded from training.

Finally, while many studies noted that URM assessments before and during training are affected by multiple social, linguistic, and cultural factors, only one group of authors attempted to measure these systematically. Two studies showed that the training performance of IMGs in the UK were associated with their linguistic and cultural understanding[40] as well as their age, sex, level of experience, and socioeconomic status.[27] The complex interaction of selection measures, selection decisions, and broader social goals is well illustrated by these studies, which conclude that existing methods intended to ensure the equivalence of doctors trained outside the UK before entering specialty training may not be achieving that purpose. The authors speculate on whether tests of IMGs English fluency in the UK might in fact be measuring other cognitive constructs, and note their results imply that it would be necessary to significantly increase the cut-offs on IMG entrance exams for those exams to actually enforce equivalence between IMGs and domestic graduates. They suggest that due to the reliance of the UK health system on IMGs, such changes would risk severe workforce shortages, and consider alternatives that balance different social goals, such as increased support for IMGs, or other methods of testing.[40]

3.5.3 - Evidence that assessments reduce specialty training diversity

Evidence on the impact of pre-training assessments on MSTS was interpreted in four main ways. Least problematic were studies which found no differences between URMs and other groups on pre-

training assessments and selection into training or later outcomes and concluded there was no evidence of barriers to diversity caused by selection methods (Table 3-3, unshaded studies). The strength of this evidence is limited by the exploratory nature of most of the studies and the absence of power analyses.

A second group of studies found evidence that the selection of URMs into medical specialties was affected by specific biases in pre-selection measures, typically because low URM pre-selection scores were not consistent with equivalent in-training performance. The evidence included gender biases affecting letters of recommendation,[41-45] sociolinguistic biases affecting selection interviews for IMGs,[27,36,46] and bias against candidates sitting the USMLE 1 including women[15,39] and IMGs.[47] This research focused on the need to measure and correct for biases, or to develop more valid alternative measures, which is also both reasonable and preliminary.

The final group of studies found that URMs had lower scores on pre-selection measures which were associated with a lower probability of selection and/or later outcomes. There were two quite different interpretations of these results. Some authors concluded that it is undesirable for low pre-selection scores to prevent URMs from entering training, even where they appear to accurately predict later performance, and suggested various ways of meliorating the impact such as relaxing cut-offs for URMs[48] or providing greater resources for IMGs.[49] Others concluded that the association of low pre-selection scores with lower scores on measures during training suggests that the under-representation is acceptable where it reflects lower levels of ability.[40,49-51]

The literature is not currently able to resolve these viewpoints. The view that URMs are under-represented because of ability rather than bias was most strongly asserted with reference to IMGs in the UK literature, while the view that pre-selection scores should not prevent URMs from entering specialty training was mainly associated with ethnicity and to a lesser extent gender in the US literature. The latter view raises the question whether there are selection methods that can facilitate

URM entry into specialty training without unacceptable tradeoffs such as significantly reduced reliability of assessments.

3.5.4 - Evidence that novel selection methods can increase training diversity

Consistent with previous reviews of the impact on diversity of medical selection methods from medical school through consultancy we found that the diversity research is focused on academic pre-selection measures such as entrance or licensure exams, due to reliability, availability, and convenience, and that there is limited evidence of selection methods likely to increase training diversity.[14,20,52] Even critics of non-specific academic pre-selection measures acknowledge that there is a need for some method of short-listing applicants for medical specialty training programs due to the highly competitive nature of a system where as many as 800 applications might be received for 5 positions on a general surgery program.[17] As a result, novel methods of selection must either replace existing reliable measures, or augment/modify them in some way.

Of two studies reporting on efforts to increase diversity of medical training by increasing the selection of URMs into training, one claimed success[38] and one did not.[37] The study claiming success did not replace the USMLE as an initial screen, but rather added a specially designed second screening tool with unreported psychometric properties. Given the main reason the USMLE 1 has been almost universally used as a specialty screen in the US is because it is highly reliable and does not require additional resources, it is unclear whether the extra resources and reduced reliability of this approach is justified by an 8% increase in URM interviews.

We did not discover any evidence suggesting that diversity can be increased by using existing measures in a different way, for example by changing the relative weight given to the various measures and methods described in Table 3-1.

3.5.5 - Lessons for global health systems

The literature provides preliminary evidence requiring replication that existing measures used for MSTS may be biased against women and IMGs in specific circumstances, and one article which showed it is possible to increase the number of URM interviews, if not the number of URMs entering training, by screening for specific characteristics. Limited reporting of statistical power leaves open the possibility that material biases against URMs exist but have not been adequately tested. Some authors concluded that the poor performance of IMGs on assessments from selection through to certification were reliable indicators of ability, although a more nuanced view was that the main issue is unequal access to cultural and linguistic resources, remediable by adequate support and training.[46]

Despite these limited results, and the absence of research outside the US and UK, the present review is relevant to other countries looking to reform their MSTS frameworks to improve diversity, particularly in the context of significant recent developments. In the US, the Federation of State Medical Boards (FSMB) and National Board of Medical Examiners (NBME) have decided to change reporting of the USMLE 1 to pass/fail rather than graded, preventing its use as a MSTS instrument;[22,23] and the University of California and other US institutions have decided to eliminate MCQ entrance exams.[24] These changes were presented as efforts to address barriers that directly contribute to the under-representation of some groups in higher education generally and medical specialist training in particular, and both highlight the relative tension between reliability and validity discussed above.[14,20] In effect, these US-based institutions have decided that the advantages of reliable assessments, which primarily benefit privileged groups, are outweighed by the disadvantages of limited validity, which tend to directly disadvantage less privileged groups, and indirectly broader society.

At the same time that use of the most common standardised MSTS instrument in the US is being prevented, the UK has moved towards greater reliance upon standardised testing, with multiple

medical colleges in the UK adopting the Multi-Specialty Recruitment Assessment (MSRA) tool.[25,26] While the evidence base is limited (for example, a PubMed search for "Multi-Specialty Recruitment Assessment" on 20.03.21 returned only 1 relevant article, a letter published in 2021), the MSRA seeks to find a better balance between reliability and validity by developing multiple sources of evidence and reducing the influence of more subjective selection methods.[27] It includes computer-based tests, including SJTs and CPSTs, which have been suggested to be relatively more valid than other measures used for medical selection.[20] It is interesting that uptake and weighting of the MSRA in selection decisions by UK medical colleges appears to have been accelerated by covid, due to the reduced social contact required by computer-based testing versus other methods like interviews.[28]

We do not propose to explore the complex broader social context which will have influenced these contrasting developments in the US and UK, other than noting the preoccupation with equity in both countries represented by movements such as Black Lives Matter[29] and #MeToo;[30] and the UK's exit from the European Union which has been linked with immigration patterns and the desire for increased quality of health care.[31] However, we suspect such factors may have played a part in the divergent paths of the US and UK with respect to MSTS, with the US relatively prioritising equity over reliability; and the UK relatively prioritising reliability while trying to improve the validity of MSTS by systematically drawing on multiple sources of evidence.

The limitations of the reviewed literature make it difficult to predict the impact of changes in MSTS frameworks intended to increase diversity. The US and UK examples suggest that other countries considering reforming their MSTS frameworks might be tempted to prioritise the reliability of pre-existing academic exams modelled on the UK, over the uncertainty associated with the US approach, however justifiable as a means of improving diversity. It is too early to judge the results of either approach. As a result, the only sure recommendation from this literature for countries hoping to improve the reliability of MSTS and increase diversity is the need to closely monitor the impact of changes to avoid or respond rapidly to unintended consequences. In the absence of evidence of

reliable selection methods that increase diversity, moving away from existing MSTS measures may leave URMs worse off,[38] particularly if specialty programs revert to methods such as alumni networks, letters of recommendation, or other techniques that are biased towards those with greater resources. While acknowledging the trade-offs between the interests of patients, minorities, and society in general, some have argued that this lack of evidence justifies selection into medical training by a weighted lottery as the only existing method likely to be effective in achieving truly equitable levels of diversity in medical workforces.[53]

Achieving increased diversity by more reliable methods than a weighted lottery will require two main advances in the literature. Current MSTS frameworks rely on pre-selection academic results rather than measures specific to specialties, alongside more subjective methods such as letters of recommendation, interview, and references. The only specialty specific measures identified in this review were for GP training (UK)[46] and a single surgical training program (US).[38] It has been argued that the use of general measures for specialty selection has led to an arms race with constantly escalating scores required for entry.[17] Developing more specific measures may allow URMs to focus on targeted knowledge and skills and to benefit from reduced competition for places. There is likely to be a trade-off between greater validity and reduced reliability for such measures given the much larger number of people who take entrance exams for medical school and licensure for medical practice than enter any medical specialty. The limited evidence available for the MSRA, adapted from the specific measures developed for GP selection,[46] makes it difficult to anticipate what impact its adoption by other medical colleges will have on the diversity of their workforces.

Second, in order to resolve whether under-representation in medical specialties is due to biased measures, differential ability, or other factors such as distribution of resources, it will be necessary to complete adequately powered prospective studies with successful and unsuccessful applicants, comparing general exam measures with specialty specific measures and accounting for the effect of confounding factors such as age, linguistic ability, cultural knowledge, and economic status. Well-

designed research should generate results that are somewhat generalisable between countries, but local conditions will always be relevant. This type of study would also help identify what support measures might be necessary to improve diversity, assuming that differential performance at the point of selection is due to unequal resources rather than differential capacity.

3.5.6 - Strengths and limitations

The review involved systematic searches of multiple databases supported by hand-search and reference-tracking, and comparison of literature from the US, UK, and Canada, with article quality evaluated using the MERSQI. It was limited by the absence of meta-analytic statistics due to the heterogeneity of the studies. Confident conclusions were limited by the exploratory nature of most of the literature, the absence of replications, and retrospective/convenience-based designs. The possibility of bias in the search strategy and/or results was explored and quantified, but cannot be entirely ruled out, although observed imbalances results were similar to a previous review with a broader focus. This is the first review to examine the impact of MSTS methods on medical workforce diversity, which is an issue of immediate interest in the context of a divergence in the US/UK use of standardised tests that may provide guidance for other countries looking to reform MSTS.

3.6 - Conclusions

Consistent with the broader medical selection literature, a focused review of the impact of MSTS methods on the diversity of medical specialist workforces suggests those actually responsible for selection decisions continue to value the reliability of pre-selection academic results, with little evidence that this is a significant cause of the under-representation of some groups, albeit the evidence base is small, underpowered, and focused almost entirely on the US and UK. Some stakeholders have prioritised alternative social goals including assessment validity and workforce

diversity. In the context of strong cultural movements addressing perceived inequities, MSTS frameworks in the US and UK are moving in different directions, with the US reducing reliance on standardised measures to promote diversity, and UK medical colleges increasing their use but attempting to improve validity by drawing on multiple sources of evidence. The fact that the two most researched MSTS frameworks are taking different paths on an uncertain evidence base demonstrates both the strong extra-scientific pressures, and the need for rigorous international longitudinal research on causes of under-representation of minorities and effective means to answer these. Countries considering MSTS reform to achieve socially accountable health systems with appropriately diverse health workforces must support systematic research in their own training systems and monitor for and respond to unanticipated consequences of change.

3.7 - References

1. Sullivan Commission on Diversity in the Healthcare Workforce. Missing Persons: Minorities in the Health Professions - A Report of the Sullivan Commission on Diversity in the Healthcare Workforce. Durham, North Carolina; 2004.

2. National Medical Training Advisory Network. Australia's Future Health Workforce – Psychiatry. Canberra, ACT; 2016.

3. Health Workforce Australia. Australia's Future Health Workforce – Doctors. 2014.

4. Shannon G, Jansen M, Williams K, et al. Gender equality in science, medicine, and global health: where are we at and why does it matter? Lancet. 2019;393(10171):560-569.

5. Department of Health. National Medical Workforce Strategy - Scoping Framework. 2019.

6. Cohen JJ, Gabriel BA, Terrell C. The case for diversity in the health care workforce. Health Aff. 2002;21(5):90-102.

7. Reeve C, Woolley T, Ross SJ, et al. The impact of socially-accountable health professional education: A systematic review of the literature. Med Teach. 2017;39(1):67-73.

8. Australian Institute of Health and Welfare. Profile of Indigenous Australians. Australia's Welfare. https://www.aihw.gov.au/reports/australias-welfare/profile-of-indigenous-australians. Published 2019. Accessed April 5, 2020.

9. Royal Australian and New Zealand College of Psychiatrists. Innovate Reconciliation Action Plan: December 2016 - December 2018. Melbourne, Victoria; 2018.

10. Toretsky C, Mutha S, Coffman J. Breaking Barriers for Underrepresented Minorities in the Health Professions.; 2018. https://healthforce.ucsf.edu/publications/breaking-barriers-underrepresented-minorities-health-professions.

11. Denney ML, Freeman A, Wakeford R. MRCGP CSA: Are the examiners biased, favouring their own by sex, ethnicity, and degree source? Br J Gen Pract. 2013;63(616):718-725.

12. Norman G. Editorial - The morality of Medical School admissions. Adv Heal Sci Educ. 2004;9(2):79-82.

13. Hecker K, Norman G. Have admissions committees considered all the evidence? Adv Heal Sci Educ. 2017;22(2):573-576.

14. Roberts C, Khanna P, Rigby L, et al. Utility of selection methods for specialist medical training: A BEME (best evidence medical education) systematic review: BEME guide no. 45. Med Teach. 2018;40(1):3-19.

15. Hewett L, Lewis M, Collins H, Gordon L. Gender Bias in Diagnostic Radiology Resident Selection, Does it Exist? Acad Radiol. 2016;23(1):101-107.

16. Buddeberg-Fischer B, Klaghofer R, Abel T, Buddeberg C. Swiss residents' speciality choices - Impact of gender, personality traits, career motivation and life goals. BMC Health Serv Res. 2006;6(137):e1-e9.

17. Bernstein J. Not the Last Word: Ending The Residency Application Arms Race—Starting with the USMLE. Clin Orthop Relat Res. 2016;474(12):2571-2576.

18. Ridgway VF. Dysfunctional Consequences of Performance Measurements. Adm Sci Q. 1956;1(2):240-247.

19. Sen Gupta T, Reeve C, Larkins S, Hays R. Producing a general practice workforce: Let's count what counts. Aust J Gen Pract. 2018;47(8):514-517.

20. Patterson F, Knight A, Dowell J, et al. How effective a selection methods in medical education? A systematic review. Med Educ. 2016;50(1):36-60

21. Cahn P. Do Health Professions Graduate Programs Increase Diversity by Not Requiring the Graduate Record Examination for Admission? J Allied Health. 2015;44(1):51-56

22. Federation of State Medical Boards, National Board of Medical Examiners. Change to pass/fail score reporting for Step 1. United States Medical Licensing Examination Website. https://www.usmle.org/incus/. Published 2020. Accessed March 17, 2020.

23. Federation of State Medical Boards (FSMB), National Board of Medical Examiners. Section 2: USMLE Step 1 and Step 2 CK Score Uses and Interpretations (Specialty Studies) Section Overview. US MLE Website. https://www.usmle.org/pdfs/incus/InCUS_Reference_List-Section2.pdf. Published 2019.

24. Nieves A. University of California eliminates SAT/ACT requirement. Politico. https://www.politico.com/states/california/story/2020/05/21/university-of-california-eliminates-sat-act-requirement-1285435. Published 2020. Accessed May 26, 2020.

25. Health Education England – Specialty Training Website – 2021 Recruitment Plans by Specialty. https://specialtytraining.hee.nhs.uk/Recruitment/2021-Recruitment-Plans-by-Specialty Accessed 19.03.21

26. Health Education England – GP Recruitment Website – Multi-Specialty Recruitment Assessment. https://gprecruitment.hee.nhs.uk/recruitment/applicant-guidance/msra Accessed 19.03.21

27. Tiffin PA, Orr J, Paton LW, Smith DT, Norcini JJ. UK nationals who received their medical degrees abroad: Selection into, and subsequent performance in postgraduate training: A national data linkage study. BMJ Open. 2018;8(7):1-16.

28. Ooi S, Ooi R. Impact of the recent changes of the Multi-Specialty Recruitment Assessment (MSRA) weightage in specialty training recruitment during the COVID-19 pandemic. Postgrad Med J 2021;Online first:e1-e2.

29. Yancy CW. Academic Medicine and Black Lives Matter: Time for Deep Listening. JAMA. 2020;324(5):435-436.

30. Malina D, Soklaridis S, Zahn C, et al. Men's Fear of Mentoring in the #MeToo Era-What's at Stake for Academic Medicine? New Eng J Med 2018;379(23):2270-2274

31. National Conversation UK – Macclesfield – Immigration, the NHS, and Brexit trade-offs. http://nationalconversation.uk/macclesfield-immigration-the-nhs-and-brexit-tradeoffs/ Accessed 19.03.21

32. Aveyard H. Doing a Literature Review in Health and Social Care: A Practical Guide. Third. Maidenhead, England: Open University Press; 2014.

33. Reed DA, Cook DA, Beckman TJ, Levine RB, Kern DE, Wright SM. Association Between Funding and Quality of Published Medical Education Research. JAMA. 2007;298(9):1002-1009.

34. Cook DA, Reed DA. Appraising the Quality of Medical Education Research Methods: The Medical Education Research Study Quality Instrument and the Newcastle-Ottawa Scale-Education. Acad Med. 2015;90(8):1067-1076.

35. De Oliveira G, Akikwala T, Kendall M, et al. Factors affecting admission to anesthesiology residency in the United States: Choosing the future of our specialty. Anesthesiol. 2012;117(2):243-251

36. Quintero AJ, Segal LS, King TS, Black KP. The personal interview: Assessing the potential for personality similarity to bias the selection of orthopaedic residents. Acad Med. 2009;84(10):1364-1372.

37. Villwock JA, Hamill CS, Sale KA, Sykes KJ. Beyond the USMLE: The STAR Algorithm for Initial Residency Applicant Screening and Interview Selection. J Surg Res. 2019;235:447-452.

38. Gardner AK, Cavanaugh KJ, Willis RE, Dunkin BJ. Can Better Selection Tools Help Us Achieve Our Diversity Goals in Postgraduate Medical Education? Comparing Use of USLME Step 1 Scores and Situational Judgment Tests at 7 Surgical Residencies. Acad Med. 2019:1.

39. Durham SR, Donaldson K, Grady MS, Benzil DL. Analysis of the 1990–2007 neurosurgery residency match: Does applicant gender affect neurosurgery match outcome? J Neurosurg. 2018;129(2):282-289.

40. Tiffin PA, Illing J, Kasim AS, McLachlan JC. Annual Review of Competence Progression (ARCP) performance of doctors who passed Professional and Linguistic Assessments Board (PLAB) tests compared with UK medical graduates: National data linkage study. BMJ. 2014;348(April):1-18.

41. Filippou P, Mahajan S, Deal A, et al. The Presence of Gender Bias in Letters of Recommendations Written for Urology Residency Applicants. Urology. 2019;134:56-61.

42. Friedman R, Fang CH, Hasbun J, et al. Use of standardized letters of recommendation for otolaryngology head and neck surgery residency and the impact of gender. Laryngoscope. 2017;127(12):2738-2745.

43. Hoffman A, Ghoubrial R, McCormick M, Matemavi P, Cusick R. Exploring the gender gap: Letters of recommendation to pediatric surgery fellowship. Am J Surg. 2019;219(6):932-936.

44. Hoffman A, Grant W, McCormick M, Jezewski E, Matemavi P, Langnas A. Gendered Differences in Letters of Recommendation for Transplant Surgery Fellowship Applicants. J Surg Educ. 2019;76(2):427-432.

45. Lin F, Oh SK, Gordon LK, Pineles SL, Rosenberg JB, Tsui I. Gender-based differences in letters of recommendation written for ophthalmology residency applicants. BMC Med Educ. 2019;19(1):1-5.

46. Patterson F, Tiffin PA, Lopes S, Zibarras L. Unpacking the dark variance of differential attainment on examinations in overseas graduates. Med Educ. 2018;52(7):736-746.

47. Lypson ML, Ross PT, Hamstra SJ, Haftel HM, Gruppen LD, Colletti LM. Evidence for Increasing Diversity in Graduate Medical Education: The Competence of Underrepresented Minority Residents Measured by an Intern Objective Structured Clinical Examination. J Grad Med Educ. 2010;2(3):354-359.

48. Edmond MB, Deschenes JL, Eckler M, Wenzel RP. Racial bias in using USMLE Step 1 scores to grant internal medicine residency interviews. Acad Med. 2001;76(12):1253-1256.

49. Esmail A, Roberts C. Academic performance of ethnic minority candidates and discrimination in the MRCGP examinations between 2010 and 2012: Analysis of data. BMJ. 2013;347(7927):1-10.

50. McManus IC, Wakeford R. PLAB and UK graduates' performance on MRCP(UK) and MRCGP examinations: Data linkage study. BMJ. 2014;348(April):1-24.

51. Wakeford R, Denney M, Ludka-Stempien K, Dacre J, McManus IC. Cross-comparison of MRCGP & MRCP(UK) in a database linkage study of 2,284 candidates taking both examinations: Assessment of validity and differential performance by ethnicity. BMC Med Educ. 2015;15(1).

52. Prober CG, Kolars JC, First LR, Melnick DE. A plea to reassess the role of United States medical licensing examination step 1 scores in residency selection. Acad Med. 2016;91(1):12-15

53. Wouters A, Croiset G, Kusurkar R Selection and lottery in medical school admissions: who gains and who loses? MedEdPublish 2018;7(4):e1-e14

Chapter 4: Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps

Authors: Andrew Amos[*1], Kyungmi Lee[2], Tarun Sen Gupta[1], and Bunmi Malau-Aduli[1,3]

* Corresponding Author

1 College of Medicine and Dentistry, James Cook University, Townsville, Australia

2 College of Science and Engineering, James Cook University, Cairns, Australia

3 School of Medicine and Public Health, University of Newcastle, Newcastle, Australia

Citation: Amos A, Lee K, Sen Gupta T, Malau-Aduli B. Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps. *Studies in Health Technology and Informatics.* 2024; 310:795-799.

This chapter describes a novel implementation of the batch form of the machine learning algorithm known as the self-organising map (SOM), the Medline database of peer-review medical literature, and how the SOM algorithm was applied to the data contained within the Medline database to create a map of the psychiatric and non-psychiatric knowledge contained within the Medline database. The chapter includes an addendum which contains a more detailed version of the manuscript that was finally accepted for publication after strict word limits significantly reduced the level of methodological detail reported.

The selection of a square rather than a hexagonal topology for the SOM is not discussed in the manuscript. As it has been established that the relative advantages and disadvantages of these two,

or alternative topologies have not been well characterised, and noting that the square topology is the dominant form in the literature, the decision to use a square topology was primarily based on pragmatic considerations. It is more convenient to visualise, the author finds it more easy to visualise, and, most importantly, the matrix form of the square topology is more suited to the organisation of the GPUs used to compute the SOM algorithm. The latter is particularly important for efforts to improve the speed of computation in order to make new functions feasible.

The software developed for this PhD, is openly available under an MIT license at: https://github.com/mongrolwarrior/MedSOM2/tree/master/MedSOM2.

4.1 - Abstract

Biases in selection, training, and continuing professional development of medical specialists arise in part from reliance upon expert judgement for the design, implementation, and management of medical education. Reducing bias in curriculum development has primarily relied upon consensus processes modelled on the Delphi technique. The application of machine learning algorithms to databases indexing peer-reviewed medical literature can extract objective evidence about the novelty, relevance, and relative importance of different areas of medical knowledge. This study reports the construction of a map of medical knowledge based on the entire corpus of the MEDLINE database indexing more than 30 million articles published in medical journals since the 19th century. Techniques used in cartography to maximise the visually intelligible differentiation between regions are applied to knowledge clusters identified by a self-organising map to show the structure of published psychiatric evidence and its relationship to non-psychiatric medical domains.

Keywords:

Machine learning; medical informatics; science of science; medical education; information science

4.2 - Introduction

Modern medical curricula are almost exclusively developed using consensus methods like the Delphi technique, which are prone to biases which have adversely affected women and other groups in various ways.[1,2] A severe lack of readily accessible and objective evidence on which to base curriculum development decisions has contributed to the problem. Visualisations of the published medical literature based on health informatics and cartographic techniques are an underutilised source of objective evidence about medical knowledge that may be capable of reducing bias.

With the exponential increase in the production of scientific knowledge, a paradigm for understanding and describing universal and domain-specific mechanisms of scientific research and progress known as the Science of Science (SciSci) has recently emerged,[3] which Skupin and colleagues extended using cartographic techniques to improve concision and intelligibility. Core to their approach are self-organising maps (SOMs), a method for mapping scientific knowledge that can project high dimensional datasets onto 2-D maps while retaining many of the topographical features of the original space.[4-6] Biuk-Aghai and colleagues show the versatility of Skupin's cartographic approach by visualising the distribution of author counts across multiple Wikipedia language editions.[7]

One of the key problems of medical curriculum development is defining the boundaries between specialist medical domains such as internal medicine, surgery, and psychiatry; and then deciding on the relative priority to be given to each in any given training program. In order to demonstrate that it is feasible to produce a SciSci map that could address this issue, the current research aimed to develop a SOM of the structure and boundaries of psychiatric knowledge within the map of broader medical knowledge represented by the published peer reviewed medical literature. The map was designed to be an objective source of evidence suitable for guiding the construction of psychiatric and general medical curricula, capable of acting as a backdrop of well-established knowledge against which emerging topics can be contrasted and understood in their full context.

4.3 - Methods

We downloaded complete sets of the MEDLINE and MeSH databases on 01.01.21 https://www.nlm.nih.gov/databases/download/pubmed_medline.html. The large dataset necessitated the use of sparse matrix representation of article data. We applied Kohonen's batch SOM algorithm, which is computationally less intensive than the original algorithm.[5] In the absence of standard methods, we experimented with different SOM sizes, starting with the 275 x 275 SOM

used by Skupin et al[4] for their much smaller dataset, looking for a plateau of topographic error to guide the final size.

After training the SOM, we applied *neuron label clustering* in which the weights used to transform inputs into each node/neuron of the trained SOM are ranked, and the largest weight defines the term dominance for that node. Adjacent nodes with the same term dominance were clustered.[4,6] Clusters were tagged as psychiatric if they included a MeSH with a psychiatric code (all codes starting with an F; for example, the MeSH "Affective Disorders, Psychotic" is coded F03.700.150). The open source platform QGIS was used to create maps using color to differentiate psychiatric and non-psychiatric clusters.

4.4 - Results

Our SOM included data from 33,375,863 MEDLINE articles and 29,917 distinct MeSH codes. We used a 350 x 350 node SOM after topographic error plateaued at 0.264 at this size. Figure 4-1 shows all term dominance clusters differentiating between those with (with a red font) and those without a psychiatric tag (black font).

Figure 4-2 expands the area of the map around the term dominance cluster "Geriatric Psychiatry", showing the association between psychiatric and medical knowledge in related areas is retained in the visualisation of the SOM. The embedding of the psychiatric MeSH *Geriatric Psychiatry* within a complex of medical MeSH is consistent with the reality of geriatric psychiatry, which combines both psychiatric and medical care and is often considered closely related to consultation liaison psychiatry for that reason.

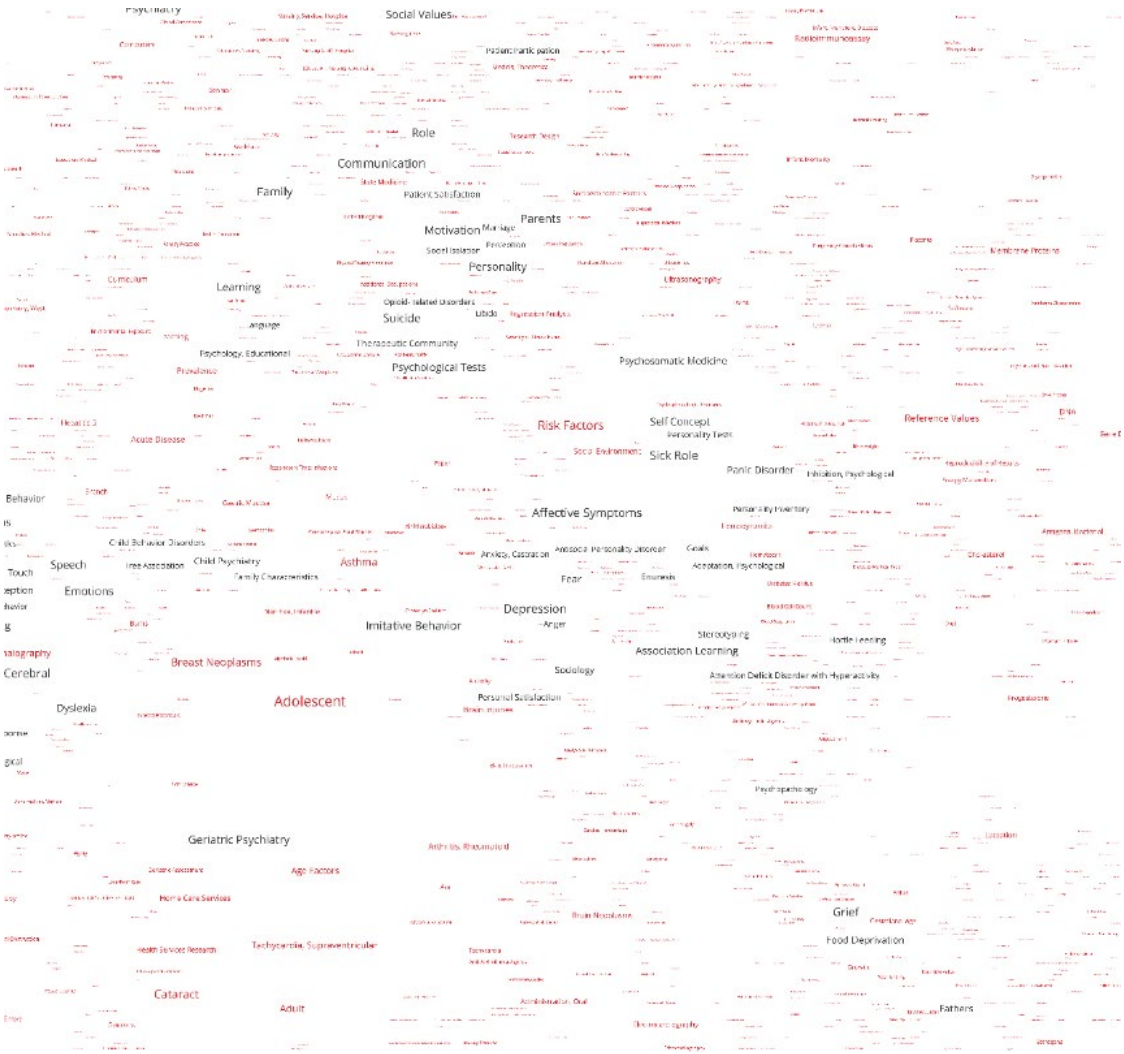


Figure 4-1. Psychiatric (black) and non-psychiatric (red) term dominance clusters (high quality image available at:

https://raw.githubusercontent.com/AndrewAmosJCU/PsychSOM/main/ColorCoded.png)

The psychiatric MeSH surrounding *Affective Symptoms* (not shown due to space constraints) include behavioural concepts such as *Sick Role*, related diagnostic groupings including *Depression* and *Panic Disorder*; related physiological states such as *Anger* and *Fear*; and predisposing factors such as *Antisocial Personality Disorder*, *Stereotyping*, and *Anxiety, Castration*. These psychiatric concepts are closely aligned with relevant medical concepts including treatment options *Benzodiazepines* and *Anticonvulsants*, with symptoms or side effects such as *Fatigue* and *Bronchoconstriction*.

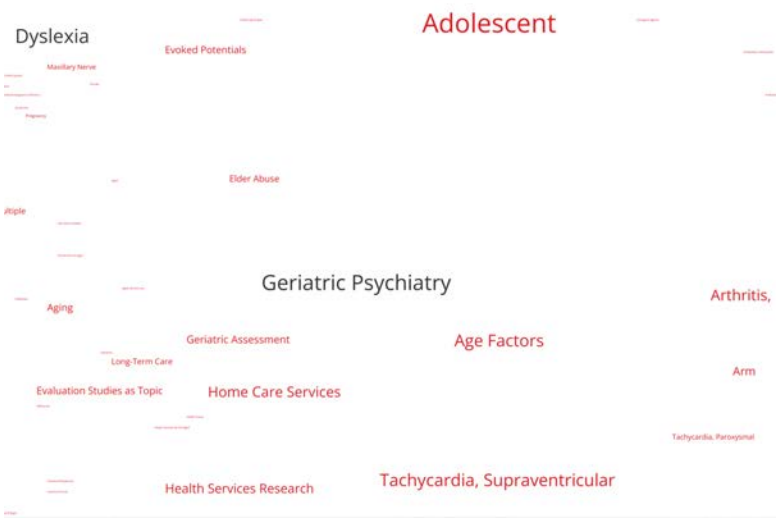


Figure 4-2. Focus on term dominance clusters around Geriatric Psychiatry.

While most of the contiguous MeSH can be readily conceptually linked in meaningful ways, some of the associations require more consideration. For example, the apposition of *Enuresis* to a region including *Fear*, *Depression*, and *Antisocial Personality Disorder* can be explained by the near neighbouring concept of *Anxiety, Castration*, which suggests the psychoanalytic link between castration anxiety and bed-wetting.

4.5 - Discussion

The results show it is possible to generate a map that summarises the entire set of articles indexed by the MEDLINE database in term dominance clusters that differentiate between psychiatric and non-psychiatric literature. For the purposes of curriculum development, the global perspective of a map derived from the most objective set of data available is less likely to be affected by systematic bias than the judgments of individual experts, or even of groups of experts from within a medical specialty or sub-specialty.

Applying cartographic techniques to SOMs provides a concise visual summary of all available peer reviewed medical evidence that can help experts understand the structure and boundaries of their own domains of expertise, and their place within the broader context of all medical knowledge. In addition to the intrinsic value of an objective standard against which to measure existing curricula, this will facilitate the identification of curricular gaps and proposed changes to address gaps or incorporate new knowledge. Future research could use this comprehensive SOM as a baseline on which to project more specific data, such as emerging topics of increasing medical research activity.

While our research has limitations characteristic of exploratory research, including using trial and error to select the size of network;[4] and accepting that the convergence properties of SOMs are not well characterised[8] we have mitigated the worst associated problems by grounding our decisions on past research and relying upon relevant quality measures such as topographic error where indicated.[9]

4.6 - Conclusions

A cartographically informed approach to SOM visualisation makes objective evidence about the structure and relationships of medical knowledge available in a condensed but readily intelligible form. As an input into medical curriculum development it has the potential to reduce historical biases that have disadvantaged women and other groups by providing a comprehensible view of the entire scope of medical knowledge.

4.7 - References

1 Thomas P, Kern DE, Hughes MT, et al. (eds). *Curriculum Development for Medical Education: A Six-Step Approach*. Third. Baltimore: Johns Hopkins University Press, 2015.

2 Sheikh MH, Chaudhary AMD, Khan AS, et al. Influences for Gender Disparity in Academic Psychiatry in the United States. *Cureus*. Epub ahead of print 22 April 2018. DOI: 10.7759/cureus.2514.

3 Fortunato S, Bergstrom CT, Börner K, et al. Science of science. *Science (1979)*; 359. Epub ahead of print 2018. DOI: 10.1126/science.aao0185.

4 Skupin A, Biberstine JR, Börner K. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. *PLoS One*; 8. Epub ahead of print 2013. DOI: 10.1371/journal.pone.0058779.

5 Kohonen T. *Self-organizing maps*. 3rd ed. Berlin: Springer, 2001.

6 Skupin A. A Cartographic Approach to Visualizing Conference Abstracts. *IEEE Comput Graph Appl* 2002; 22: 50–58.

7 RP Biuk-Aghai, CI Pang and YW Si. 2014. Visualising Large-scale Human Collaboration in Wikipedia.Future Generation Computer Systems, 31, pp. 120-133, Elsevier. doi:10.1016/j.future.2013.04.001

8 Cheng Y. Convergence and Ordering of Kohonen's Batch Map. *Neural Comput* 1997; 9: 1667–1676.

9 Pölzlbauer G. Survey and Comparison of Quality Measures for Self-Organizing Maps. In: *Proc. 5th Workshop Data Analysis, Slovakia, 2004*. Slovakia, http://www.cis.hut.fi/projects/somtoolbox (2004).

4.8 - ADDENDUM: Mapping the terrain of psychiatric knowledge within medicine with health informatics and cartography

Authors: Andrew Amos[*1], Kyungmi Lee[2], Tarun Sen Gupta[2], and Bunmi Malau-Aduli[2,3]

* Corresponding Author

1 College of Medicine and Dentistry, James Cook University, Townsville, Australia

2 College of Science and Engineering, James Cook University, Cairns, Australia

3 School of Medicine and Public Health, University of Newcastle, Newcastle, Australia

This addendum is the original, longer form of the shorter manuscript published in *Studies in Health Technology and Informatics* in 2024 and reproduced here as Chapter 4 above. This addendum provides a more detailed description of the methodology used to construct the self-organising map and discussion of the results.

4.8.1 - Abstract.

Biases in selection, training, and continuing professional development of medical specialists arise in part from reliance upon expert judgement for the design, implementation, and management of medical education. Reducing bias in curriculum development has primarily relied upon consensus processes modelled on the Delphi technique. The application of machine learning algorithms to databases indexing peer-reviewed medical literature such as MEDLINE allows for the extraction of objective evidence about the novelty, relevance, and relative importance of different areas of medical knowledge. This study reports the construction of a map of medical knowledge based on the entire corpus of the MEDLINE database indexing more than 30 million articles published in medical journals since the 19th century. Techniques used in cartography to maximise the visually intelligible differentiation between regions are applied to knowledge clusters identified by a self-organising map to show the structure of published psychiatric evidence and its relationship to non-psychiatric medical domains.

Keywords:

Machine learning; medical informatics; science of science; medical education; information science

4.8.2 - Introduction

In the twentieth century, approaches to medical education were determined by the large, rapidly changing, and highly differentiated knowledge sets required for medical specialist practice, apprentice models of training, and hierarchical work environments.[1] Medical curricula were and continue to be almost exclusively developed by consensus methods such as the Delphi technique.[2] While consensus methods can reduce biases attributable to individuals, they may be less effective in reducing group biases given the significant imbalances which persist at more senior levels of specialist practice. For example, 75% of the Directors of Training responsible for the curriculum that

produces psychiatrists via the Royal Australian and New Zealand College of Psychiatrists training program are male.[3]

Gender biases in medicine have been associated with a broad range of problems, from the erroneous assumption that heart disease is largely similar in women and men,[4] to reduced selection of women and other minority groups into medical training based on inaccurate evaluation of their aptitude for the work.[1,5] Given that medical academic preferment and decision-making responsibility may be more associated with dominant personality traits than competence,[6] alternatives to consensus-based methods for curriculum development are necessary to address biases associated with gender and other characteristics.

The heavy reliance upon expert judgement subject to consensus methods for medical curriculum development[7] means that the selection of content for inclusion is subject to the biases discussed above. Visualisations of the published medical literature based on health informatics and cartographical techniques can provide an objective source of evidence about the structure, novelty, and relative importance of psychiatric knowledge at different levels of specificity and describe how it relates to other medical domains.

4.8.2.1 - Science of Science: Mapping the mechanisms of knowledge

With the exponential increase in the production of scientific knowledge, a new paradigm for understanding and describing universal and domain-specific mechanisms of scientific research and progress has developed known as the Science of Science (SciSci).[8] Machine learning algorithms based on the analytic capacities of neural networks can extract meaningful patterns from large sets of data.[9] A key idea is that the processes that produce knowledge, as well as the knowledge itself, can be understood as networks of nodes representing scientists, ideas, journals, or other physical or intellectual constructs, joined by edges representing different types of relationship, such as co-

authorship on a research paper, supporting/refuting evidence, or similarities/differences between entities such as articles in a conceptual space such as the set of published peer reviewed literature.

4.8.2.2 - Visualisation and cartography

There are a number of machine learning approaches that transform information about the set of peer-reviewed medical research contained in large databases like MEDLINE, Scopus, or Web of Science into a network representing abstract knowledge.[10] Visualising the conceptual similarities and differences between scientific articles on two-dimensional maps generally involves dense networks where similarity is indicated by proximity and color. For example, Figure 4A-1 shows relationships between countries in the production of articles on "Medical Health Informatics" between 2020-2022, created using VOSviewer from Web of Science data.[11] The size of each node represents the number of articles produced in that country, the size of the edges between nodes represents the number of articles with an author in both countries, and patterns of frequent collaboration between countries are indicated by color.

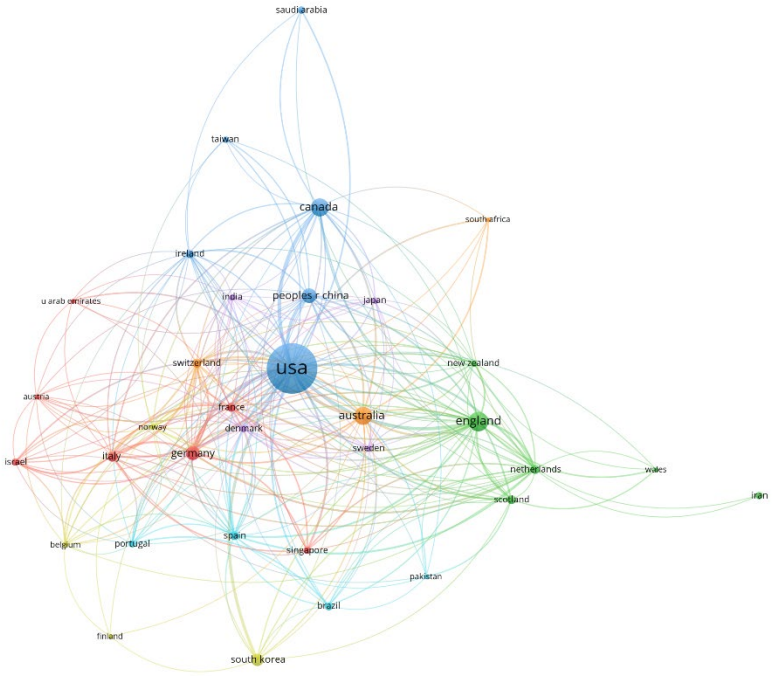


Figure 4A-1: Visualisation of Medical Health Informatics articles by country using VOSviewer.

Skupin and colleagues have extended this approach using principles developed within cartography to visually differentiate physical, political, or social regions and the relationships between them.[12] They noted that visually intelligible maps of the physical and political characteristics of countries or other geographic entities must filter out irrelevant detail, highlight important features, and frequently show overlapping sets of information in the same visual space. So, for example, just as a map of a geographic region might prioritise the labelling of countries over towns and superimpose lines representing borders overlapping colors representing mountains and rivers, a map of scientific knowledge might prioritise the labelling of fields of study such as psychiatry over subspecialties like forensic psychiatry and superimpose lines differentiating fields of study overlapping colors representing the level of activity of research in an area.

4.8.2.3 - Representing scientific literature with self-organising maps

Many techniques are used to visualise information about scientific literature.[10] Self-organising maps (SOMs), pioneered by Kohonen, are an attractive method that can be applied to structured or unstructured data, can identify novel patterns not anticipated by researchers, and have successfully projected very high dimensional datasets onto 2-D maps while retaining many of the topographical features of the original space.[12,13] These features are suited to eliciting, summarising, and compactly representing the properties of scientific knowledge latent in the published literature as an objective source of evidence suitable for inclusion in the development of medical curricula.

Elsewhere we have reported research on identifying emerging topics of psychiatric research that should be considered for inclusion in curriculum development or renewal.[14] Effective integration of emerging topics into a medical curriculum requires understanding of the context from within which the topics have emerged, and would be facilitated by tools that situate the emerging topics within more established knowledge structures. The current research aims to develop a map of the structure and boundaries of psychiatric knowledge within the map of broader medical knowledge represented by the published peer reviewed medical literature. The map is designed to be an objective source of evidence suitable for guiding the construction of medical curricula, capable of acting as a backdrop of well-established knowledge against which emerging topics can be contrasted and understood in their full context.

4.8.3 - Methods

4.8.3.1 - Datasets

While Web of Science and Scopus are more complete databases of scientific literature, we selected the MEDLINE database for our research, for multiple reasons. The MEDLINE database is the most objective and complete set of information about medical knowledge freely available in the public

domain. This database records and makes publicly available the title, authors, publication source, and publication date of articles published in most peer-reviewed medical journals since the 19th century, among a rich set of other data.[15] It is specifically devoted to the published medical literature; it is commonly used by medical and allied health professionals to identify clinically relevant research; and the entire database can be accessed without charge, which is likely to be important for many curriculum development efforts, particularly in developing countries.

To improve the utility of the MEDLINE database, the National Library of Medicine in the US maintains a controlled vocabulary of ~30,000 Medical Subject Headings (MeSH), which are phrases with precise meanings used to describe the content of medical articles across many categories, such as population, disease, treatment, and methodological approach.[15] The MEDLINE database describes the content of every indexed article with a list of MeSH of variable length, which is most often used to find specific information during database searches but is also useful for estimating how similar or different articles or groups of articles are. We downloaded the MEDLINE https://www.nlm.nih.gov/databases/download/pubmed_medline.html and MeSH https://www.nlm.nih.gov/databases/download/mesh.html databases on 01.01.21. For every article we extracted the MeSH list which describes the content and methods used.

4.8.3.2 - Data transformation

The large dataset necessitated the use of sparse matrix representation of the article data. Rather than a feature vector with a boolean variable indicating the presence or absence of each MeSH, we represented each article as a set of indices recording only the features present in that article, reducing the size of the input data by several orders of magnitude (Table 4-1).

Table 4-1. Conversion of feature vector to sparse vector representing an individual article (first 7 MeSH shown)

MeSH	Mental disorder	Geriatric psychiatry	Neurocognitive disorders	Physician-patient relations	Risk-taking	Grief	Human-animal bond	Smoking	...	➡
Present	True	False	False	False	False	True	False	True	...	...
Index	0	1	2	3	4	5	6	7	...	...
Sparse vector	0	5	7							

4.8.3.3 - Training the dataset

Due to the large size of the dataset we applied Kohonen's batch SOM algorithm, which is computationally less intensive than the original algorithm (Algorithm 1).[13] In the absence of a standard method for selecting the size and shape of the SOM we experimented with different sizes, starting with the 275 x 275 SOM used by Skupin et al[12] for their much smaller dataset as a baseline. We incrementally increased the size of both dimensions of the SOM by 25 until increasing the size did not improve the quality of the map, as measured by topographic error.[13]

Algorithm 1: Batch Sparse SOM[16]

Input x: Article set - sparse matrix with ~33 million rows, each representing 1 article with a set of indices indicating which MeSH are present (see Table 4-1) (long)

Data w: SOM codebook - 350 x 350 nodes x 29,917 weights per node (real)

Data χ: $\chi_i = \sum_j x_{ij}^2$ - array containing sum of squares of weight j for article i

Data d: Array containing distance from article i to best matching node

Data b: Array containing best matching node for each article i

Data n: Array of numerator values accumulated per epoch to calculate new weights

---

for all articles i calculate $\chi_i = \sum_j x_{ij}^2$ and initialise array **χ**

---

for all training epochs e:

calculate σ - neighborhood

for all articles i set distance d_i to ∞

for all nodes k find best matching unit

ω = $\sum_j w_{kj}^2$

for all articles i

calculate distance = ω + χ_i - 2($x_i \cdot w_k$)

if distance < d_i then store new bmu b_i and new distance d_i

for all nodes k

set den(ominator) to 0

for all weights d **set** num(erator) to 0

for all articles i accumulate den and num

c = bmu_i

h = $\frac{\exp(\|r_k - r_c\|^2)}{2\sigma^2}$

den = den + h

for all weights j

$num_j = num_j + hx_{ij}$

update all weights j of node k

$\mathbf{w_{kj}} = num_j/den$

4.8.3.4 - Visualisation

We extended Skupin et al's cartographic approach to visualising scientific information represented by large SOMs.[12,17] After training the SOM, we applied neuron label clustering in which the weights

used to transform inputs into each node/neuron of the trained SOM are ranked from highest to lowest, and the highest ranked weight defines the term dominance for that node. Adjacent nodes with the same term dominance were then clustered together. Finally, term dominant clusters were tagged as psychiatric in nature if they included a MeSH with a psychiatric code (known as an F-code because they are represented by a code starting with an F; for example, the MeSH "Affective Disorders, Psychotic" is coded F03.700.150).

We used the open source Geographic Information System (GIS) program QGIS and the open source C++ package Matplot++ to visualise the SOM trained on the MEDLINE database. The square matrix of nodes comprising the SOM were divided into distinct non-overlapping term dominant clusters. Matplot++ was used to generate an overview showing the complete set of term dominant clusters represented as polygons assigned random colors.

QGIS was used to generate a map representing the term dominant clusters from psychiatric and non-psychiatric literature. Each cluster was represented by its dominant MeSH, with the font size determined by the number of nodes contained within the term dominant cluster. A logarithmic transformation was required to prevent large clusters from dominating small clusters. Two versions of the map were produced: the first included only psychiatric MeSH, and the second included all MeSH where the MeSH label was colored black if it represented an F-code (psychiatric) tag, and red if it did not (non-psychiatric).

4.8.4 - Results

As we were not analysing the title/abstract or other language-specific data we included the entire set of 33,375,863 MEDLINE articles and 29,917 distinct MeSH codes downloaded on 1/1/2021 in our analyses. We did not exclude articles based on language, methodology, date of publication, or other criteria.

Experimenting with different sizes of SOM starting at 275 x 275 demonstrated that topographic error plateaued at close to 0.264 with a 350 x 350 lattice, with no improvement up to 400 x 400. As the risk of overfitting increases with increased size, we used the 350 x 350 SOM for our analyses.

Figure 4A-2 shows the overview of all term dominance clusters created using MatPlot++ and assigning random colors to each of the 7428 term dominant clusters. Figure 4A-3 shows all term dominance clusters with a psychiatric tag. Figure 4A-4 shows all term dominance clusters differentiating between those with (with a red font) and those without a psychiatric tag (black font).

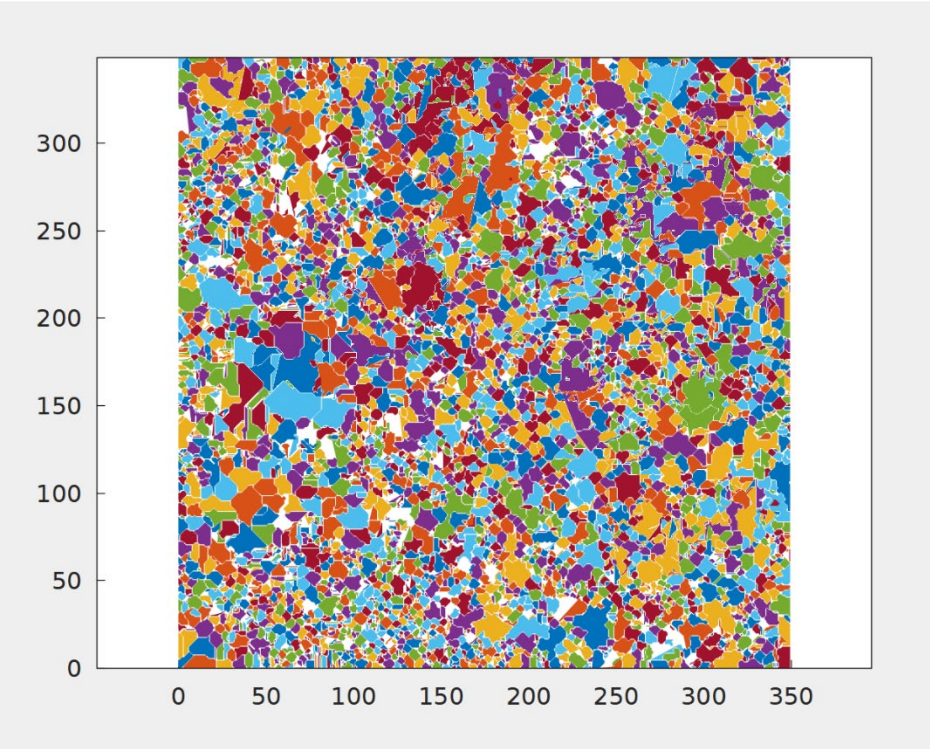


Figure 4A-2: Overview of 7428 term dominance clusters across 350 x 350 nodes of the Self-Organizing Map.

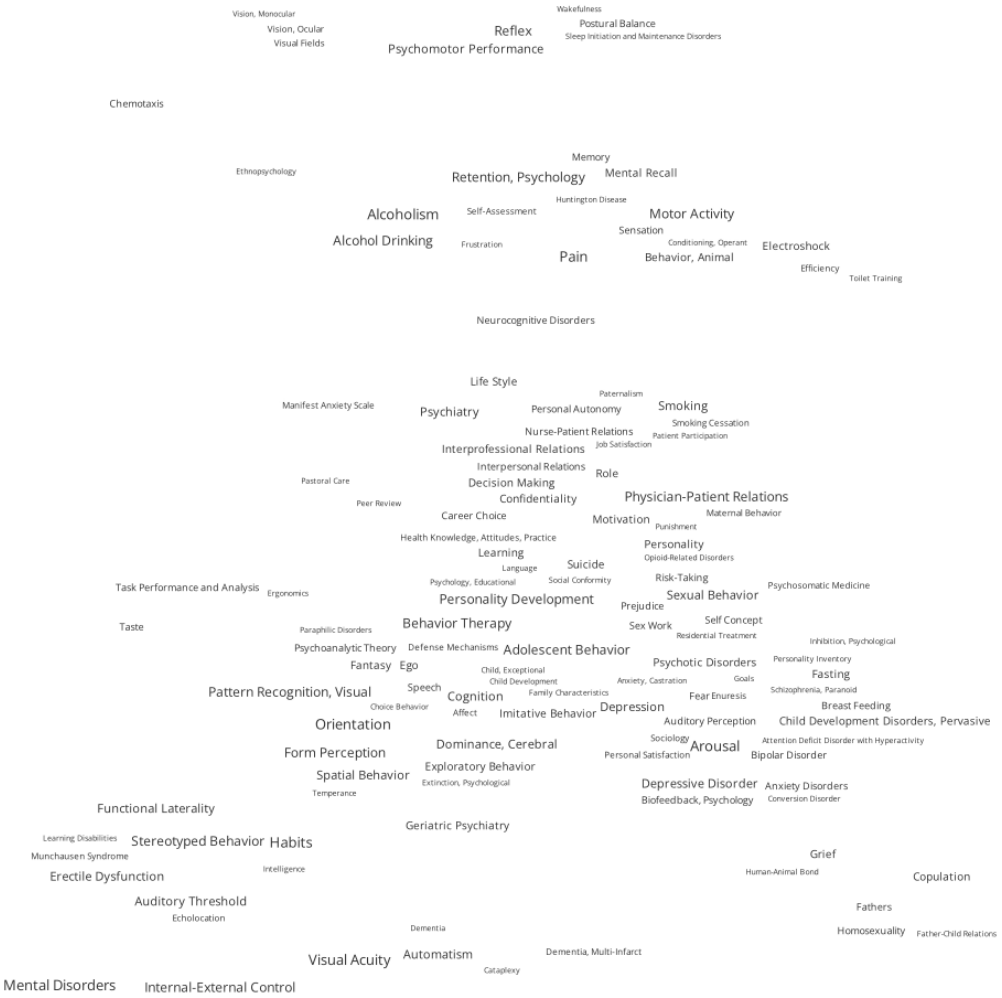


Figure 4A-3: Term dominance clusters with a psychiatric tag.

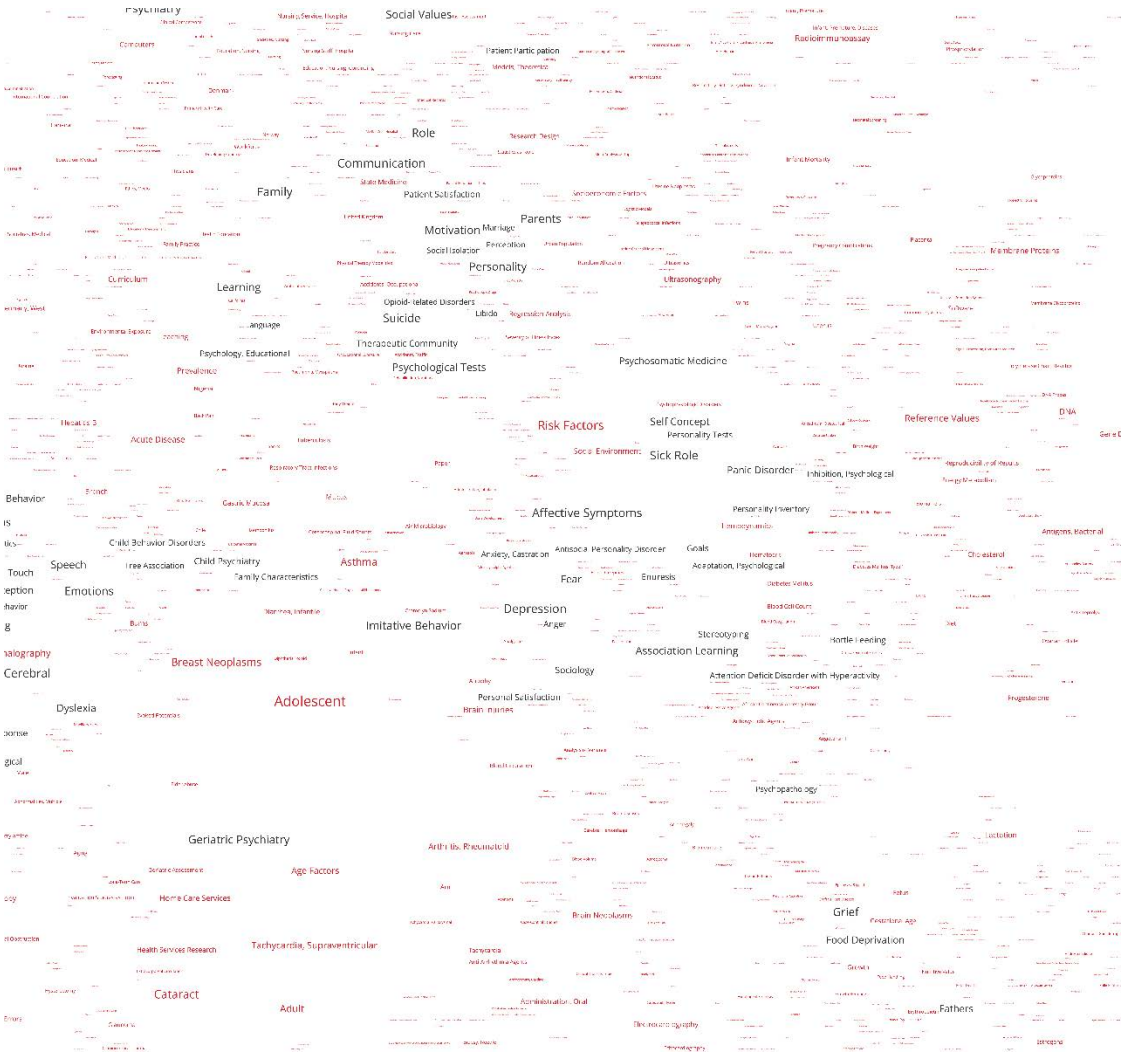


Figure 4A-4: Psychiatric (black) and non-psychiatric (red) term dominance clusters.

Figure 4A-5 expands the area of the map around the term dominance cluster "Geriatric Psychiatry", showing the association between psychiatric and medical knowledge in related areas is retained in the visualisation of the SOM.

The embedding of the psychiatric MeSH Geriatric Psychiatry within a complex of medical MeSH (Figure 4A-5) is consistent with the reality of geriatric psychiatry, which combines both psychiatric and medical care and is often considered closely related to consultation liaison psychiatry for that reason.

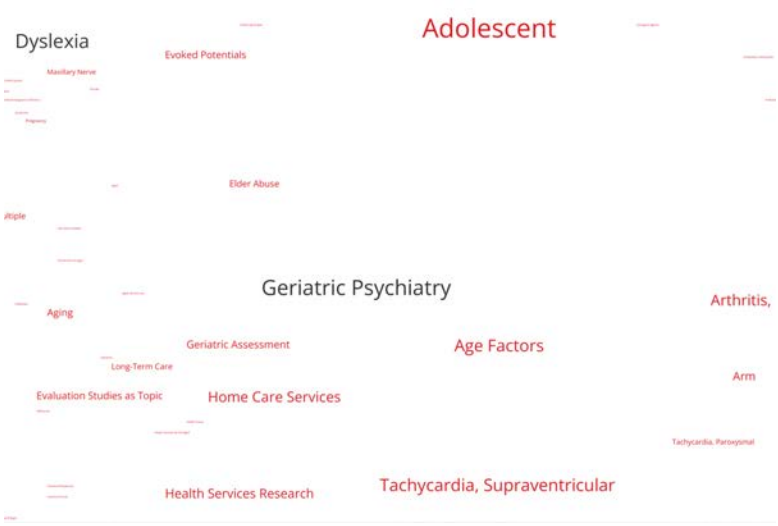


Figure 4A-5: Focus on term dominance clusters around Geriatric Psychiatry.

The psychiatric MeSH surrounding Affective Symptoms (Figure 4A-6) include behavioural concepts such as Sick Role, related diagnostic groupings including Depression and Panic Disorder; related physiological states such as Anger and Fear; and predisposing factors such as Antisocial Personality Disorder, Stereotyping, and Anxiety, Castration. These psychiatric concepts are closely aligned with relevant medical concepts including treatment options Benzodiazepines and Anticonvulsants, with symptoms or side effects such as Fatigue and Bronchoconstriction.

While most of the contiguous MeSH can be readily conceptually linked in meaningful ways, some of the associations require more consideration. For example, the apposition of Enuresis to a region including Fear, Depression, and Antisocial Personality Disorder can be explained by the near neighbouring concept of Anxiety, Castration, which suggests the psychoanalytic link between castration anxiety and bed-wetting.

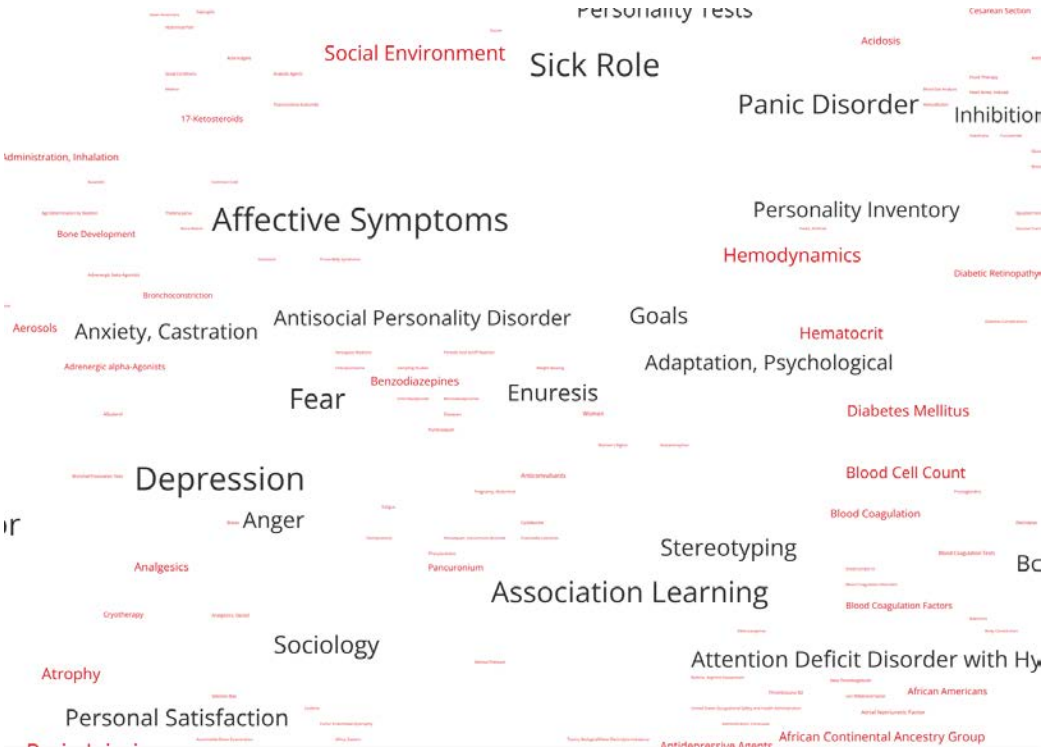


Figure 4A-6: Focus on term dominance clusters around Affective Symptoms.

4.8.5 - Discussion

The results demonstrate that it is possible to generate a map that summarises the entire set of articles indexed by the MEDLINE database in term dominance clusters that differentiate between psychiatric and non-psychiatric literature. Inevitably a map summarising information about more than 30 million articles by projecting 29,917 features onto a two-dimensional map involves significant abstraction, but there are clear indications of the retention of higher-order meaning in the lower dimensional relationships.

4.8.5.1 - The Map of Science as a canvas for curriculum development

Our work demonstrates that it is possible to extract an outline of the complete set of published peer reviewed medical research indexed by MEDLINE, and to map the boundaries between psychiatric and non-psychiatric concepts within that conceptual space. One of the advantages of grounding a

map of science in the peer reviewed literature is that it facilitates an understanding of the high-level organisation of scientific knowledge which can be directly linked to more specific knowledge by drilling down into individual topics, authors, or articles. Chen and Song have reported a cascading citations method which shows how to integrate global, abstract levels of knowledge with detailed, local articles.[18]

4.8.5.2 - Integration of objective data with expert judgement in curriculum development

This feature of SciSci maps may be particularly useful for addressing biases in the medical curriculum development process. The global perspective of a map derived from the most objective set of data available is less likely to be affected by the biases of individual experts, or even of a group of experts from within a medical specialty or sub-specialty.

It is not currently feasible to replace the central role of expert judgement in medical curriculum development. As a result, to make use of SciSci maps in curriculum development will require the development of techniques for integrating this source of information into the curriculum development process. Further studies could be based on work by Chen and colleagues.[19] This group showed that self-organising maps can be used to understand the conceptual organisation of topics discussed by content experts. The SOMs rapidly extracted many of the themes identified by the experts themselves, but also extracted relevant concepts implicit in the expert discussion, but not consciously identified by those experts. This suggests that information provided by SOMs can be useful to the process of curriculum development carried out by experts in several ways:

- Rapid mapping of main topics in literature and discussion (explicit knowledge)
- Clarification of relationships across domains
- Identification of implicit knowledge

Thus, applying cartographic techniques to the presentation of SOMs provides a concise visual summary of all available peer reviewed medical evidence that can help experts understand the structure and boundaries of their own domains of expertise, and their place within the broader context of all medical knowledge. In addition to the intrinsic value of an objective standard against which to measure existing curricula, this will facilitate the identification of curricular gaps and proposed changes to address gaps or incorporate new knowledge. Future research could use this comprehensive SOM as a baseline on which to project more specific data, such as emerging topics of increasing medical research activity.

4.8.5.3 - Limitations

The self-organising maps reported here are designed to extract patterns from unstructured data and do not have well-established quality measures or criteria for selecting parameters such as map size or number of training epochs. As a result, some of the design decisions rely upon trial and error, for example by varying the map size by 25 neurons in both dimensions over a range. In addition, the SOM batch method relies upon convergence across multiple training epochs, which has only been proven for one dimensional maps, and not the two dimensional maps used here.[13,20] While these are not necessarily unexpected limitations of exploratory research, we have tried to avoid the worst associated problems by grounding our decisions on past research (such as the size of the map[12]) and relying upon relevant quality measures such as topographic error, where indicated.[21]

4.8.6 - Conclusion

The sheer volume of medical knowledge, the organisation of medical workforce into specialist streams, and an apprenticeship approach to medical education have all contributed to the reliance upon expert judgement for prioritising different areas of medical knowledge for research, practice, and education. Reliance upon expert judgement has some advantages, but also some problems, particularly possible biases tending to disadvantage less represented groups.

Machine learning techniques such as the self-organising maps reported in this study, now provide the opportunity to develop objective sources of evidence which may both improve the efficiency of medical curriculum development and reduce biases in the selection of content to be studied during medical training and professional development.

Capitalising on this potential will require development of methods for integrating these new forms of information into the consensus processes currently used for medical curriculum development. We propose that cartographic methods, designed to take advantage of highly developed human visual systems using cues including color, size, and proximity to represent overlapping sets of information, are ideal for this purpose.

4.8.7 - References

1 Carraccio CL, Englander R. From Flexner to competencies: Reflections on a decade and the journey ahead. Academic Medicine 2013; 88: 1067–1073.

2 Thomas P, Kern DE, Hughes MT, et al. (eds). Curriculum Development for Medical Education: A Six-Step Approach. Third. Baltimore: Johns Hopkins University Press, 2015.

3 Royal Australian and New Zealand College of Psychiatrists. Discussion Paper: Gender equity and the College: Why does it matter? Melbourne, https://www.ranzcp.org/files/membership/wellbeing/ranzcp-discussion-paper-gender-equity.aspx (2022, accessed 16 November 2022).

4 The Lancet. Cardiology's problem women. The Lancet 2019; 393: 959.

5 Amos AJ, Lee K, Sen Gupta T, et al. Systematic review of specialist selection methods with implications for diversity in the medical workforce. BMC Med Educ; 21. Epub ahead of print 1 December 2021. DOI: 10.1186/s12909-021-02685-w.

6 Sheikh MH, Chaudhary AMD, Khan AS, et al. Influences for Gender Disparity in Academic Psychiatry in the United States. Cureus. Epub ahead of print 22 April 2018. DOI: 10.7759/cureus.2514.

7 Grant J. Principles of Curriculum Design. In: Swanwick T, Forrest K, O'Brien B (eds) Understanding Medical Education: evidence, theory, and practice. Oxford, UK: Wiley Blackwell, 2019, pp. 71–88.

8 Fortunato S, Bergstrom CT, Börner K, et al. Science of science. Science (1979); 359. Epub ahead of print 2018. DOI: 10.1126/science.aao0185.

9 Brunn M, Diefenbacher A, Courtet P, et al. The Future is Knocking: How Artificial Intelligence Will Fundamentally Change Psychiatry. Academic Psychiatry 2020; 44: 461–466.

10 Boyack KW, Klavans R. Creation and Analysis of Large-Scale Bibliometric Networks. In: Glänzel W, Moed HF, Schmoch U, et al. (eds) Springer Handbook of Science and Technology Indicators. Cham, Switzerland: Springer Nature, 2019, pp. 187–212.

11 van Eck NJ, Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics 2017; 111: 1053–1070.

12 Skupin A, Biberstine JR, Börner K. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS One; 8. Epub ahead of print 2013. DOI: 10.1371/journal.pone.0058779.

13 Kohonen T. Self-organizing maps. 3rd ed. Berlin: Springer, 2001.

14 Amos AJ, Lee K, Sen Gupta T, et al. Identifying emerging topics in the psychiatric literature to facilitate curriculum renewal and development. Current Psychology; Submitted.

15 Mork JG, Yepes AJJ, Aronson AR. The NLM medical text indexer system for indexing biomedical literature. CEUR Workshop Proceedings 2013; 1–13.

16 Melka J, Mariage J. Efficient Implementation of Self-Organizing Map for Sparse Input Data. In: Proceedings of the 9th International Joint Conference on Computational Intelligence. Funcha, Madeira, Portugal, 2017, pp. 54–63.

17 Skupin A. A Cartographic Approach to Visualizing Conference Abstracts. IEEE Comput Graph Appl 2002; 22: 50–58.

18 Chen C, Song M. Visualizing a field of research: A methodology of systematic scientometric reviews. PLoS One; 14. Epub ahead of print 1 October 2019. DOI: 10.1371/journal.pone.0223994.

19 Chen H, Schuffels C, Orwig RE. Internet Categorization and Search: A Self-Organizing Approach. J Vis Commun Image Represent 1996; 7: 88–102.

20 Cheng Y. Convergence and Ordering of Kohonen's Batch Map. Neural Comput 1997; 9: 1667–1676.

21 Pölzlbauer G. Survey and Comparison of Quality Measures for Self-Organizing Maps. In: Proc. 5th Workshop Data Analysis, Slovakia, 2004. Slovakia, http://www.cis.hut.fi/projects/somtoolbox (2004).

Chapter 5: Identifying emerging topics in the peer-reviewed literature to facilitate curriculum renewal and development

Authors:

Andrew James Amos[*1], MB.BS, Kyungmi Lee[2], PhD, Tarun Sen Gupta[1], PhD, Bunmi S. Malau-Aduli[1,3], PhD

* Corresponding Author

1 College of Medicine and Dentistry, James Cook University, Townsville, Australia

2 College of Science and Engineering, James Cook University, Cairns, Australia

3 School of Medicine and Public Health, University of Newcastle, Newcastle, Australia

This chapter describes the application of a bibliometric technique which analyses the relative incremental increase in the number of articles published on each topic of current medical interest in a medical literature database to identify the topics with the greatest recent increase in research activity relative to their previous baseline.

5.1 - Abstract

Objective

This article reports a bibliometric analysis of emerging topics in the psychiatric literature indexed in the MEDLINE database as a technique for renewal of clinical training curricula.

Methods

Summary data of English-language articles indexed in the MEDLINE database between 1971-2018 were downloaded. Emerging topics in nine demi-decades between 1972-1976 and 2012-2016 were identified by the incremental incidence of individual Medical Subject Headings (MeSH) compared with previous years. Co-word analysis was used to investigate and visualise the relationships between emerging topics in each demi-decade.

Results

Summaries of 18 million articles annotated with psychiatric/psychological MeSH were retrieved and used to identify emerging topics. Peaks in the number of articles annotated by the top 20 emerging topics in 9 demi-decades coincided with release of the third and fourth editions of the Diagnostic and Statistical Manual which codifies psychiatric diagnoses. Themes emerging from network visualisations of the most common emerging MeSH in each demi-decade were consistent with movements in psychiatric/psychological theory and practice since the 1970s, including the recent focus on psychological and social factors implicated in suicide and suicide prevention.

Conclusions

The identification of emerging topics within the published medical literature is a viable technique for use in curriculum renewal projects as a counterweight to biases driven by expert judgement. While indices like MEDLINE make the published literature an appealing initial step in building an empirical

basis for curriculum development, it also demonstrates the potential value of less public and less structured data, such as health service electronic medical records.

Keywords:

Trends in life science; Emerging topics; MeSH; PubMed; Curriculum development; Medical education

5.2 - INTRODUCTION

Modern clinical training curricula seek to promote excellent patient care by helping junior health workers acquire and maintain effective and up-to-date skills, knowledge, and attitudes. While there are well-established and systematic approaches to clinical curriculum development, due to the enormous, rapidly growing amount of health research, and the complex, unbounded, and rapidly changing nature of health work, these approaches rely heavily on expert judgement (Harden, 2001; P. Thomas et al., 2015). Unfortunately, the unconscious biases of health experts are now recognised to have had negative effects across many areas of health care. The most prominent example may be the exclusion of female patients from medical trials on the incorrect assumption that they would show the same patterns of illness and treatment response as male patients, leading to many years of suboptimal treatment of cardiovascular disease in women (The Lancet, 2019). Our own work has examined how the unconscious biases of experts may contribute to the under-representation of minority groups in specialist medical training (Amos et al., 2021; Roberts et al., 2018).

Existing curriculum development approaches attempt to reduce bias almost exclusively by expert consensus, using methods such as the Delphi technique, which assumes that a synthesis of the beliefs of a diverse group of stakeholders will be less subjective than the beliefs of any individual stakeholder (Thomas et al., 2015). While the peer reviewed literature has long been the most objective source of evidence about health, illness, and treatment, only recently has the emergence of data mining techniques made it possible to extract high-level objective evidence about the

nature, structure, and relative importance covering all aspects of health care from large databases such as hospital admissions, medication prescriptions, and health outcomes (Brunn et al., 2020). The current article describes an effort to apply data mining techniques to databases describing the peer reviewed medical literature to provide objective evidence suitable for use in curriculum development to help reduce the unconscious biases which are likely to arise from expert judgement.

5.2.1 - Curriculum development and expert bias

Although frameworks such as those developed by Harden (Harden, 2001) and Kern (Thomas et al., 2015) highlight the need for periodic revision of medical education curricula to add the most important recent innovations and remove outdated information, there is no standard method to identify which of the many changes in clinical knowledge should be considered, or how to judge the relative importance of different areas of practice (Benson et al., 2019). As a result, other than expert consensus, there are no standard methods for identifying or correcting biases in health curricula. Existing approaches either do not address curricula content selection at all (Swanwick et al., 2018), or focus on topic- or process-specific questions, such as describing how a small group of experts in a local training program might refine their expertise with reference to models of learning (Kulasegaram et al., 2018) or feedback from stakeholders (Benson et al., 2019). Rather than using empirical evidence to identify or reduce systemic sources of bias, research has focused on mechanisms leading to bias in individual stakeholders (Risberg et al., 2009), or proposed that there is an ethical imperative to reform clinical education, without exploring how reform can be achieved, or how it would improve training or patient outcomes (Braun & Saunders, 2017). Even research which explicitly analyses the biases inherent in expert-driven curricula makes limited use of empirical evidence to identify gaps in coverage or over-represented material (D'Eon & Crawford, 2005). Compared with the narrow focus of these methods, efforts to identify and correct expert bias using

data mining can rely on objective evidence drawn from large databases that cover all areas of health and health care.

5.2.2 - Bibliometric analysis

Among many other applications, data mining techniques have been used to identify high-level patterns in science such as the most active areas of scientific research (Zitt et al., 2019). Scientometrics, which analyses quantitative features of science, such as the size, growth, and relationships between different forms of scientific knowledge, makes heavy use of data mining. Within scientometrics, bibliometrics focuses on the quantitative features of scientific communication such as peer-reviewed publications, including network maps that visualise underlying implicit knowledge structures (van Raan, 2019).

The uses of network maps in bibliometric analysis are rapidly evolving, but can be usefully summarised in terms of goals, data sources, and techniques. At a general level, bibliometrics uses network maps to understand and explain the structure and relationships of diverse phenomena such as research communities and domains of knowledge (Boyack & Klavans, 2019). For example, network maps have been used to visualise concrete instantiations of the scientific research paradigms proposed by Kuhn to understand the influence of social dynamics on scientific activity (Chen, 2003; Kuhn, 1962). While data mining techniques work on any structured or unstructured data set, bibliometric analyses often use structured databases containing summaries of published peer review literature because of their technical advantages, including reduced computational load and increased interpretability. Despite requiring commercial licenses for large-scale access, Web of Science (WoS) and Scopus are the two most prominent scientific research indexes because of the detailed information they collect across all or most scientific research and related materials. MEDLINE is frequently used for the analysis of published medical literature because it is free to use, well structured, and extensively used by clinicians (Boyack & Klavans, 2019).

The bibliometric techniques used to map scientific constructs involve four main steps: extracting data from primary sources (e.g. collating search results from a database such as MEDLINE); calculating similarities between entities (for example, how similar is the research output of two researchers, or two universities); clustering (essentially grouping relatively similar entities together and separating relatively different entities); and visualisation (converting abstract information into meaningful visual patterns) (Boyack & Klavans, 2019). Four main alternatives are used to generate the linkages constituting bibliometric networks: co-citation, co-word, co-author, and bibliographic coupling. Co-citation analyses link entities by the number of times they have both been cited by a third entity (paradigmatically two peer reviewed articles both cited by a third article). Co-word analyses link entities by the frequency with which they use the same words (for example, two articles which use a large number of the same words in their title and abstracts are linked in this way). Co-author analyses examine networks of authors or other actors who have been authors on the same articles. Finally, bibliographic coupling networks link publications if they both cite a third publication (the reciprocal of co-citation analysis) (Moral-Munoz et al., 2019).

5.2.3 - Emerging topics

Recently, bibliometric researchers have developed techniques for analysing the large sets of published articles indexed in databases such as MEDLINE to identify the most active areas of research and debate, described by one set of authors as emerging topics (Ohniwa et al., 2010; Ohniwa & Hibino, 2019; Wang et al., 2018). While it cannot replace expert judgement in the curriculum management process, automated identification of emerging topics has the potential to mitigate existing biases by providing a more objective and rapid basis for identifying the topics to be considered for inclusion in training curricula. In addition to the phrase “emerging topics”, researchers in this area have described many alternative techniques using diverse terms including

"emerging trends" (Chen, 2006), "research fronts" (Åström, 2007; Persson, 1994), "scientific revolutions" (de Langhe, 2017), and "innovation" (Vernon et al., 2021), among others.

There does not appear to be a dominant paradigm for the investigation of what we will describe as emerging topics hereafter. The literature reports the use of co-word analyses (Persson, 1994; Wang et al., 2018), co-citation analyses (Åström, 2007; Chen, 2006), and bespoke quantitative analyses of dynamic changes in publication patterns over time (Ohniwa et al., 2010; Ohniwa & Hibino, 2019). Bibliographic coupling and co-author analyses seem less often used to identify emerging topics, which for the former may be due to the time lag before two articles can acquire a common citing source.

5.2.4 - MEDLINE and Medical Subject Headings

MEDLINE is the most accessible large index of peer-reviewed medical literature. It records the title, abstract, and authors of every article published in almost all well-established peer-reviewed medical journals for more than a century, alongside a wealth of other information. The National Library of Medicine (NLM), which maintains MEDLINE, also annotates every indexed article with Medical Subject Headings (MeSH) that classify the patients, area of specialty, research methods, and other characteristics. Ohniwa et al used these MeSH to identify emerging topics across the whole of medicine by calculating which MeSH were starting to be used much more often than in previous years, which they described as the increment rate (Ohniwa et al., 2010; Ohniwa & Hibino, 2019).

The MeSH used to describe the topics addressed by each article indexed in the MEDLINE database are an example of a controlled vocabulary. Demotic languages such as English are uncontrolled in the sense that while there are formal frameworks such as dictionaries and grammatical guidelines which suggest meaning and structure, for most purposes the rules are not enforced. Controlled vocabularies, by contrast, establish precise meanings and specific rules for use. The meanings and

usage of MeSH are established and maintained by the NLM using a combination of semi-automated bibliometric analyses interpreted and applied by human experts (with primary expertise in classification rather than necessarily in medical knowledge domains) (Mork et al., 2013). This is an ideal model for developing techniques by which bibliometric analyses might support curriculum development by human experts, as it demonstrates the complementary abilities of computers, which can rapidly analyse, organise, and present summaries of enormous sets of data; and humans, who are much better at disambiguating specific usages of polysemous words, identifying exceptions, making attributions regarding relative importance, and contextualising research, for example by understanding of social context.

5.2.5 - Bibliometric analysis and curriculum development

Our research was designed to explore how bibliometric data mining techniques might be used to support curriculum development and maintenance by identifying emerging topics from the published medical literature in a way that would be independent of the biases of individual experts. While it is likely that the published literature itself contains biases, as a primary source of evidence about medical knowledge and practice it is difficult to conceive of a more complete and less biased source of evidence. In addition, unlike the biases associated with individual experts, it is possible for the biases within the published literature to be systematically identified and addressed. In the absence of a gold standard method for the identification of emerging topics within peer-reviewed literature, we selected the model used by Ohniwa et al (2019) as the most intuitively understandable, and the most widely accessible due to its use of the publicly available MEDLINE database rather than the more comprehensive but less specific and commercially restricted WoS and Scopus databases.

We expected that the pattern of emerging topics in psychiatry and psychology would be influenced by the release of new editions of the Diagnostic and Statistical Manual of Mental Disorders (DSM) in

1980, 1994, and 2013 (American Psychiatric Association, 1980, 1994, 2013) due to the significant research required to revise this handbook of psychiatric diagnosis. We hypothesized that applying the methods described by Ohniwa et al (2010) to the published psychiatric literature would generate an evidence base that could be used to guide psychiatric curriculum development independent of expert judgement. Our aim was to describe the emerging topics in the psychiatric literature since 1972, 8 years before the third edition of the DSM in 1980, to show how they could be integrated into the curriculum development framework and examine the potential of data mining other sources for curriculum development and renewal.

5.3 - Materials and methods

5.3.1 - Datasets

While Web of Science and Scopus are the most complete and most commonly used datasets for bibliometric analyses, we selected the MEDLINE database for our research, for two main reasons. MEDLINE is a well-structured database maintained by the US-based NLM that covers a large corpus (defined as a set of documents) of the scientific research most relevant to our study, with an established controlled vocabulary constructed by human experts aided by automated bibliometric data analysis (Mork et al., 2013). Perhaps more importantly, as we intended to develop a technique able to be used for curriculum development across all settings, MEDLINE is publicly available, while WoS and Scopus require commercial licenses for access to the level of data required for the type of analysis developed here.

We downloaded the complete MEDLINE (https://www.nlm.nih.gov/databases/download/pubmed_medline.html) and MeSH (https://www.nlm.nih.gov/databases/download/mesh.html) databases on 01.01.21. Although we planned to look at the years 1972 – 2016, calculation of the increment statistic described below requires one year of leading and two years of lagging data. After selecting articles published in

English between 1971 and 2018, summary data about 18,072,356 articles described by 29,640 medical subject headings (MeSH) were included in the analyses. Each article is labelled ("annotated" in the bibliometric literature) with a variable number of MeSH, where each MeSH represents a medical subject that is present in the article. For example, an article about the treatment of bipolar affective disorder with the medication lithium would be annotated with MeSH for "Bipolar disorder" and "Lithium" as well as others specifying the type of clinical trial, the population, and so on. Following Ohniwa et al, we extracted only unique MeSH tags and excluded terms unrelated to research topics (excluding the top 2 levels of the MeSH tree and MeSH terms under categories M, N, V, and Z) (Ohniwa et al., 2010; Ohniwa & Hibino, 2019). The MeSH hierarchy groups all psychiatry and psychology terms under a common branch – all psychiatry and psychology MeSH have codes beginning with the letter F. To consider only articles with psychiatric content, we excluded all indexed articles which did not contain at least one F-coded MeSH. Of the 29,640 MeSH there were 1123 unique MeSH under the "Psychiatry and Psychology" main heading beginning with the letter F included in our study.

5.3.2 - Emerging topics

Following Ohniwa et al, the equation for the increment rate (*I*) of MeSH term α in year β was:

$$I_{\alpha \text{ in } \beta} = X_{\alpha \text{ in } \beta} / Y_{\alpha \text{ in } \beta}$$

where $X_{\alpha \text{ in } \beta}$ = total appearances of α in years $\beta + 1$ and $\beta + 2$

and $Y_{\alpha \text{ in } \beta}$ = total appearances of α in years $\beta - 1$, β, $\beta + 1$ and $\beta + 2$

Emerging topics were defined as MeSH in the top 5% of $I_{\alpha \text{ in } \beta}$ of each year β (Ohniwa et al., 2010).

The increment rate *I* can be thought of as a ratio comparing the number of appearances of a MeSH in two consecutive years with its appearances in those years plus the two previous years. An

emerging topic will have a small number of appearances in the two prior years compared with the two subsequent years, leading to a relatively large *I* close to 1.0. This will also force the *I* of previously emerging topics that have reached a plateau towards 0.5. The context for emerging topics was examined using the accumulation of articles annotated with all psychiatry/psychology MeSH over the period of study and the number of articles annotated with emerging topics alone.

5.3.3 - Co-word analysis

Co-word analysis measures how similar the information content of individual documents within a corpus is to other documents by how many words with similar meanings co-occur in each document. It relies upon the assumption that the words used in each document represent ideas important in understanding scientific background, methodology, or evidentiary claims, and that co-occurrence of the same words in different documents means that they address the same ideas (Callon et al., 1983). Our co-word analysis was performed using the MeSH annotating each article (considered as words within a controlled vocabulary), rather than the plain text words of the titles, abstracts, or document bodies. First, we created a list of emerging topics comprising the MeSH with the top 5% of *I* rates in each year. From this list of MeSHs, the 20 that appeared most frequently in each five-year period (demi-decade) were selected and analysed with the statistical package R (R Core Team, 2020) to discover the extent to which they annotated the same articles. To simplify the visualisations, only MeSH contained in the top 20 were represented by vertices, and only the top 5% most common co-word occurrences were shown as edges.

5.3.4 - Application of emerging topics to psychiatric curricula

To illustrate how the emerging topics may be applied to curriculum development, we analysed the most recent World Psychiatric Association curriculum published in 2002 (World Psychiatric Association, 2002) with reference to the top 5 keywords in each of the three demi-decades.

5.4 - Results and Discussion

5.4.1 - Emerging topic analysis

At the time of extraction the MEDLINE database indexed close to 30 million articles of which around 18 million were annotated with at least one MeSH indicating a link with psychiatry/psychology (Figure 5-1). Figure 5-2 shows the number of articles annotated with emerging topics alone, per period. Consistent with our expectation, Figure 5-2 shows emerging topic peaks associated in time with publication of different editions of the DSM, discussed below. Putative emerging topics in psychiatry are represented by Table 5-1, which shows the top 20 most frequently appearing emerging topics in each period. While it might be expected that each topic would only appear once on the list of emerging topics, with a period of rapidly increasing frequency of use followed by a plateau, a small number appeared on two or three separate occasions, indicating a number of peaks followed by regression. This might be the result of the same topic appearing at different times associated with different innovations (for example, a single medication being associated with different conditions), or peaks and troughs of interest in the same topic (for example, a treatment with initial promise being discontinued due to side effects, but later revisited when the side effects can be managed). Table 5-2 shows which MeSH appeared as emerging topics in more than one period.

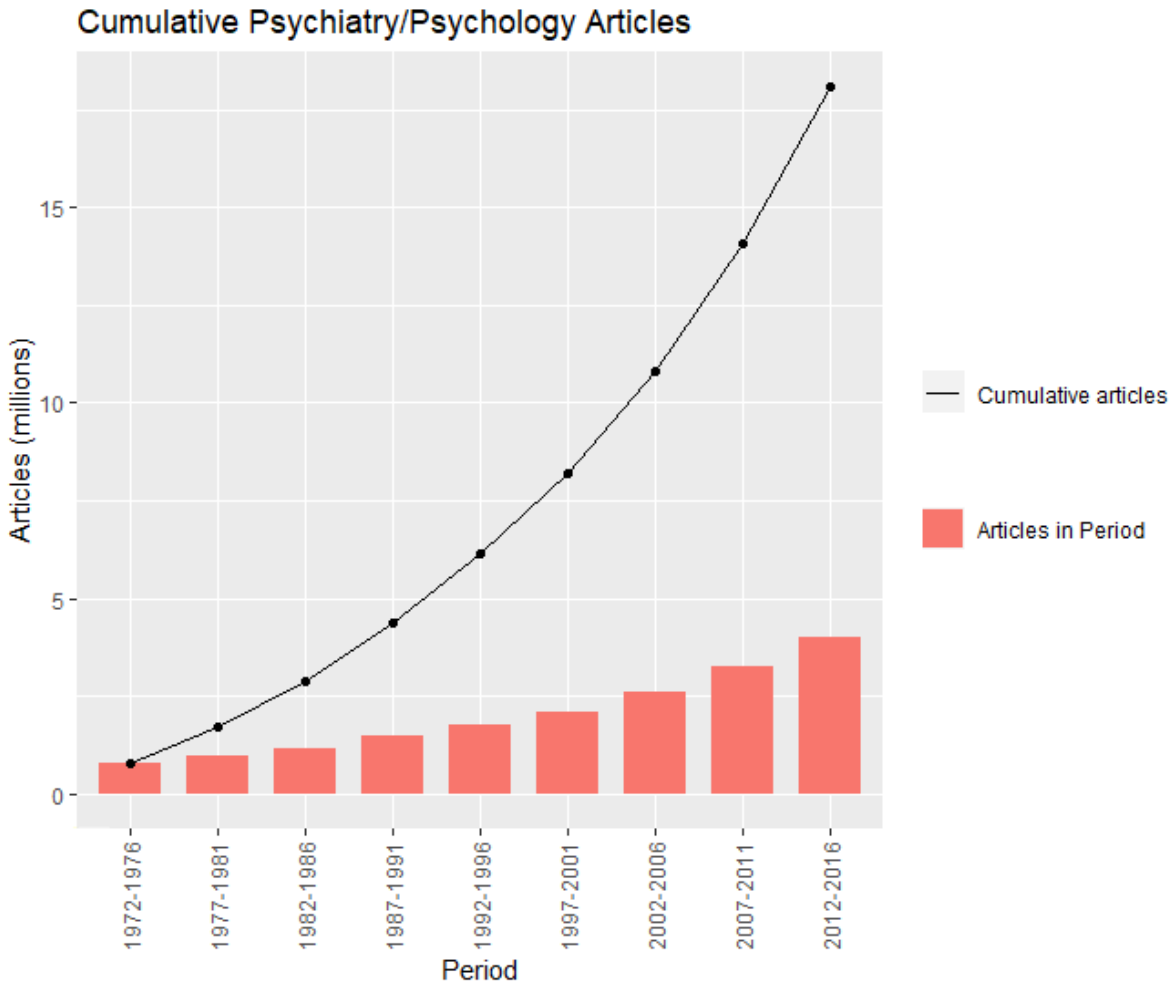


Figure 5-1: Cumulative articles annotated with psychiatry/psychology MeSH (1972-2016)

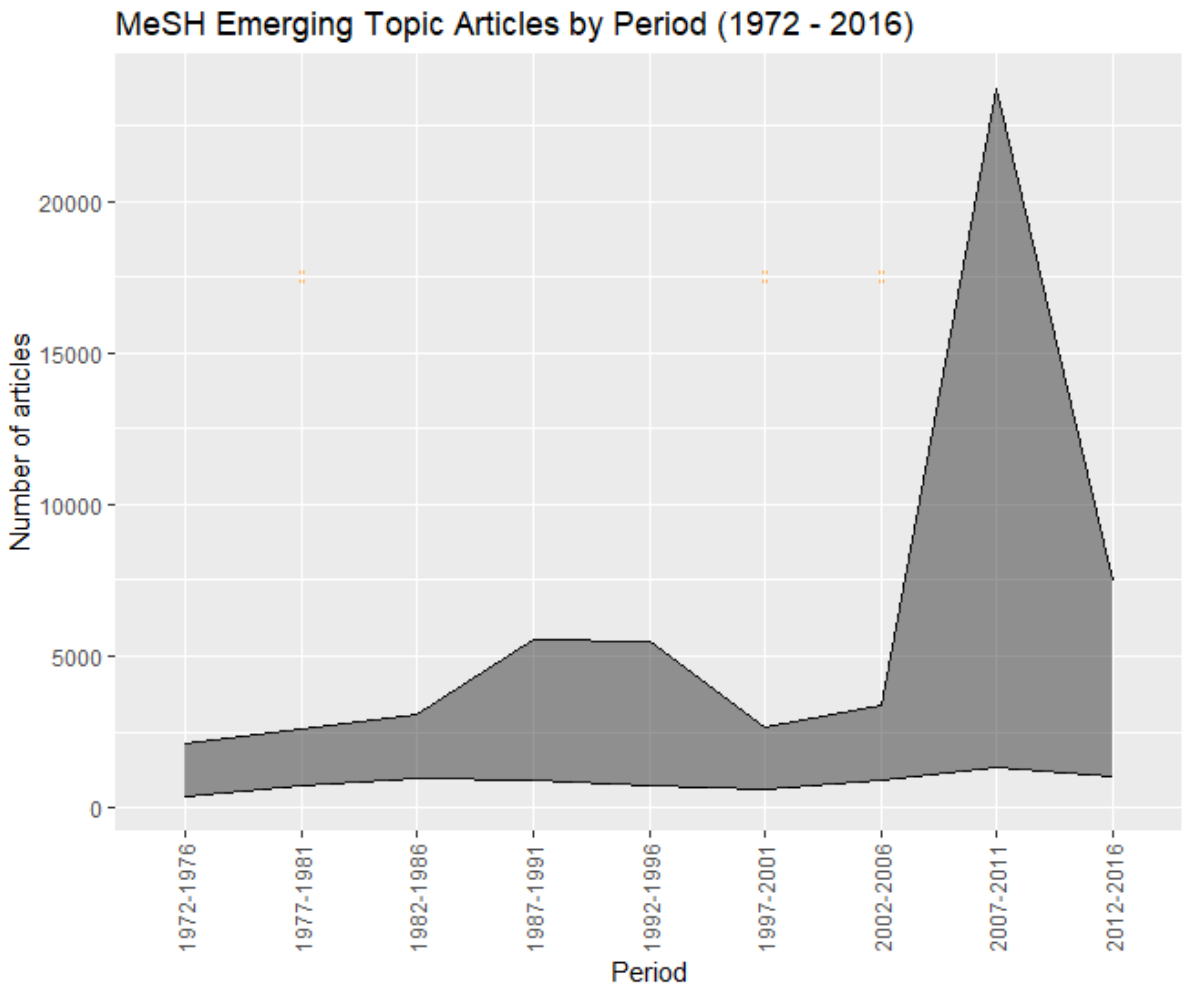


Figure 5-2: Number of articles annotated with emerging topics per period

Table 5-1: Top 20 most frequently appearing emerging topics in each demi-decade

1972-1976		1977-1981		1982-1986		1987-1991		1992-1996		1997-2001		2002-2006		2007-2011		2011-2016	
Family	2123	Physician-Patient Relations	2605	Attitude of Health Personnel	3056	Attention	5530	Health Knowledge, Attitudes, Practice	5511	Cues	2666	Personality Inventory	3386	Neuropsychological Tests	23688	Animal Distribution	7496
Discrimination Learning	1248	Dominance, Cerebral	2288	Adaptation, Psychological	2950	Arousal	4948	Patient Satisfaction	5296	Cooperative Behavior	2385	Recognition, Psychology	2727	Psychiatric Status Rating Scales	14147	Cognitive Dysfunction	5633
Nurse-Patient Relations	1062	Patient Compliance	1829	Arousal	2305	Psychiatric Status Rating Scales	4072	Personality Assessment	3352	Task Performance and Analysis	2043	Social Perception	2671	Risk Reduction Behavior	3590	Sedentary Behavior	4502
Consumer Behavior	967	Discrimination Learning	1816	Attention	2221	Neuropsychological Tests	3390	Internal-External Control	2722	Erectile Dysfunction	1573	Memory, Short-Term	2507	Medication Adherence	3416	Consensus	3745
Judgment	863	Bipolar Disorder	1502	Homosexuality	2160	Mental Recall	2874	Mental Health Services	1620	Depressive Disorder, Major	1498	Psychophysics	2496	Speech	3068	Social Stigma	3685
Chemotaxis	857	Pattern Recognition, Visual	1268	Alzheimer Disease	1974	Personality Development	2471	Suicide, Assisted	1216	Disclosure	1417	Empathy	2345	Interview, Psychological	3056	Opioid-Related Disorders	3295

Psychometrics	836	Arousal	1244	Psychomotor Performance	1904	Discrimination Learning	2289	Empathy	1141	Personal Autonomy	1299	Siblings	1939	Psychological Tests	2980	Thinking	3284
Identification, Psychological	684	Psychiatric Status Rating Scales	1200	Mental Recall	1802	Orientation	2266	Cues	1126	Mental Processes	1244	Association Learning	1664	Child Behavior	2778	Suicidal Ideation	3230
Mental Health Services	663	Mental Recall	1124	Learning	1660	Personality Tests	1929	Social Perception	1075	Cocaine-Related Disorders	1225	Narration	1640	Uncertainty	2710	Tobacco Use Disorder	2973
Dominance, Cerebral	645	Dementia	1103	Consumer Behavior	1586	Problem Solving	1908	Group Processes	1047	Nurse's Role	1176	Trust	1461	Social Identification	2534	Autism Spectrum Disorder	2835
Stereotyped Behavior	636	Depressive Disorder	998	Nurse-Patient Relations	1497	Personality Inventory	1890	Parenting	963	Psychological Theory	1057	Interdisciplinary Communication	1444	Executive Function	2282	Resilience, Psychological	2367
Schizophrenic Psychology	624	Affective Symptoms	998	Suicide	1475	Leadership	1445	Erectile Dysfunction	905	Judgment	1035	Inhibition, Psychological	1401	Consensus	1796	Burnout, Professional	2287
Learning Disabilities	582	Learning Disabilities	980	Physician's Role	1449	Internal-External Control	1428	Homosexuality, Male	902	Self Efficacy	930	Risk Reduction Behavior	1309	Language Tests	1644	Bullying	2032
Attitude to Death	573	Speech Perception	958	Vision, Ocular	1269	Cognition Disorders	1349	Psychology, Child	884	Social Identification	827	Concept Formation	1295	Arthralgia	1560	Marijuana Smoking	1667

Self-Assessment	568	Social Responsibility	954	Orientation	1263	Health Behavior	1327	Pain Threshold	830	Tobacco Use Disorder	795	Comprehension	1191	Efficiency	1555	Gambling	1594
Goals	507	Life Change Events	848	Fear	1187	Health Knowledge, Attitudes, Practice	1313	Self-Assessment	779	Problem-Based Learning	724	Esthetics	1133	Child Development Disorders, Pervasive	1526	Mindfulness	1357
Efficiency	479	Sleep Stages	833	Neurocognitive Disorders	1177	Patient Participation	1113	Maze Learning	775	Peer Review, Research	694	Achievement	1046	Impulsive Behavior	1526	Touch Perception	1299
Psychophysics	465	Discrimination, Psychological	825	Confidentiality	1149	Language	1106	Nuclear Family	748	Inhibition, Psychological	684	Uncertainty	977	Ergonomics	1453	Bisexuality	1275
Individuality	417	Job Satisfaction	824	Anorexia Nervosa	951	Individuality	1014	Achievement	744	Facial Expression	656	Intention	974	Psycholinguistics	1419	Racism	1136
Object Attachment	392	Psychometrics	745	Job Satisfaction	951	Substance Abuse, Intravenous	927	AIDS Dementia Complex	729	Dementia, Vascular	629	Language Tests	892	Sexuality	1326	Spatial Memory	1053

Table 5-2: Top 20 emerging topics appearing in more than one demi-decade

Frequency	Emerging topic	
Appears three times	Arousal	
	Discrimination Learning	
	Mental Recall	
	Psychiatric Status Rating Scales	
Appears twice	Achievement	Mental Health Services
	Attention	Language Tests
	Consensus	Learning Disabilities
	Consumer Behavior	Neuropsychological Tests
	Cues	Nurse-Patient Relations
	Dominance, Cerebral	Orientation
	Efficiency	Personality Inventory
	Empathy	Psychometrics
	Erectile Dysfunction	Psychophysics
	Health Knowledge, Attitudes, Practice	Risk Reduction Behavior
	Individuality	Self-Assessment
		Social Identification

Inhibition, Psychological	Social Perception
Internal-External Control	Tobacco Use Disorder
Job Satisfaction	Uncertainty
Judgment	

A separate network showing the number of annotated articles for each emerging topic and the strength of relationship between each keyword is reported for each demi-decade between 1972 and 2016 (Figure 5-3). The size of each node represents the number of articles containing each emerging topic MeSH, while the thickness of the line between each pair of nodes represents how frequently different MeSH appeared in the same articles. This illustrates how the identified emerging topics in some demi-decades are arranged around one core concept, such as the “Family” node for 1972-1976, while other demi-decades feature a number of relatively disconnected sub-networks, such as 1977-81 with distinct clusters around “Bipolar disorder”, “Cerebral dominance”, and “Physician-patient relations”.

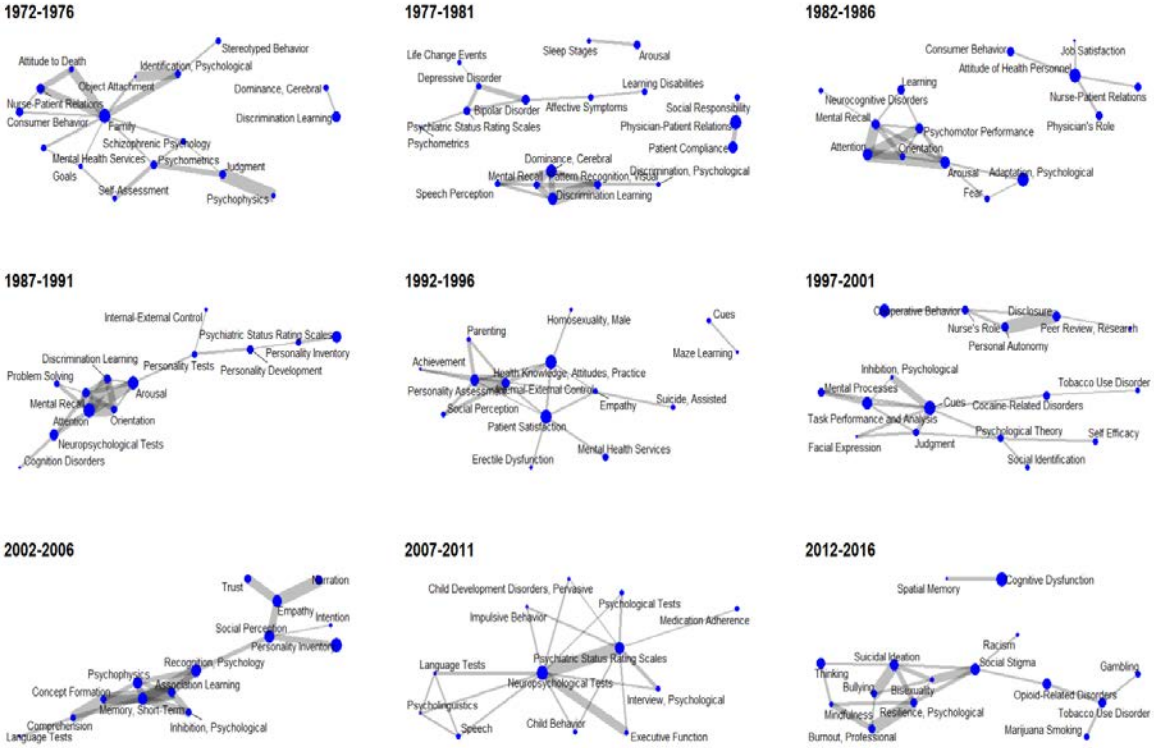


Figure 5-3: Emerging topic networks (1972 – 2016)

Table 5-3 shows examples of how emerging topics could be used to identify areas of the WPA curriculum and how the WPA syllabus could be reviewed with reference to the latest literature. For example, emerging topic activity annotated with the “Personality Inventory” MeSH would trigger attention to the WPA curriculum items “Didactic Curriculum” and “Syllabus”.

Table 5-3: Cues for curriculum/syllabus renewal based on emerging topics

Period	Top 5 Keywords in Period	Curriculum/syllabus review
2002-2006	Personality Inventory Recognition, Psychology Social Perception Memory, Short-Term Psychophysics	Didactic curriculum (WPA p9-10) • Human growth and development • Basic knowledge – classification Syllabus (WPA 54+) • Growth and development • Diagnostic instruments
2007-2011	Neuropsychological Tests Psychiatric Status Rating Scales Risk Reduction Behavior Medication Adherence Speech	Didactic curriculum • Neurosciences • Adult psychopathology Syllabus • Neurology/Neuropsychiatry • Psychopharmacology
2012-2016	Animal Distribution Cognitive Dysfunction Sedentary Behavior Consensus Social Stigma	Didactic curriculum • Neurosciences and biological risk factors • Geriatric/old age psychopathology Syllabus • Psychiatric rehabilitation • Sexual/gender issues

The bibliometric approach proved able to condense the information contained in the indexed summaries of more than 18 million published medical articles and extract useful information about the most cited emerging topics across four decades. Figure 5-2 shows two peaks of emerging topics associated with the top 20 MeSH in a period, with a doubling of activity between 1987 – 1996, and an order of magnitude increase between 2007 – 2011. We speculate that these periods were associated with a concentration of activity triggered by the creation and publication of two editions of the Diagnostic and Statistical Manual (DSM; 4th and 5th editions in 1994 and 2013, respectively (American Psychiatric Association, 1994, 2013)). This interpretation is supported by the presence of MeSH "Psychiatric Status Rating Scales" and "Neuropsychological Tests" in the top 5 most annotated lists of both periods. "Psychiatric Status Rating Scales" was also in the top 10 most annotated list for the period 1977-1981, overlapping the publication of the 3rd Edition of the DSM in 1980 (American Psychiatric Association, 1980).

Several patterns emerge from visualisations of the most annotated emerging topics and their relationships in each period (Figure 5-3). Over all periods there was a greater than expected tendency for all 20 top emerging topics to form relatively coherent networks. The second of the nine periods had four separate clusters of activity (1977-1981), with five periods having two clusters (including three with one dominant and one relatively minor cluster) and three completely connected networks (1972-6, 1987-1991, and 2002-2006).

Dominant themes of each period are outlined in Table 5-4.

Table 5-4 shows that the themes emerging out of this analysis are consistent with the trends observable during the periods reviewed. Our primary interest is how this information can be used in the curriculum development process. To illustrate how this might work, we have used the curriculum of the World Psychiatric Association, published in 2002, and indicated how the results of each period since then might

have been used to identify areas of the curriculum and syllabus which would require reconciliation with the emerging literature. We have taken examples both from the didactic curriculum (essentially an overview at an abstract level of the domains of knowledge, skills, and attitudes required for competent psychiatric practice) and the syllabi (the specific learning experiences planned to acquire the features outlined in the curriculum (World Psychiatric Association, 2002).

Table 5-4: Themes emerging in each period

Period	Keywords Curriculum/syllabus review
1972-76	The most common MeSH was Family, forming a nexus between psychological MeSH including Object Attachment, service-oriented terms like Nurse-Patient Relations and Consumer Behavior, and a clinical subnetwork comprising Goals, Psychometrics, and Self-Assessment. This combination of emerging topics reflects two transitions occurring at the time: from a more biological model of psychiatry focused on severe mental illnesses like schizophrenia to a more bio-psycho-social approach including more prevalent conditions like anxiety and depression; and from a psychoanalytic tradition (Object Attachment) towards psychological approaches (Psychology, Psychometrics, and Self-Assessment).
1977-81	The only period with four disconnected clusters shows the influence of the DSM-III published in 1980, with a cluster dominated by formulation of the mood disorder category comprising Psychiatric Status Rating Scales, Psychometrics, Bipolar Disorder, and Depressive Disorder. There is a neuroscience cluster including

Discrimination Learning, Psychological Discrimination, Mental Recall, and Visual Pattern Recognition. Finally, there are two smaller clusters with a biological group linking arousal and Sleep Stages, and a clinical services group containing Social Responsibility, Physician-Patient Relations, and Patient Compliance.

1982-86 The larger of two clusters shows the growing importance of neuropsychology (as distinct from neuroscience) linking cognitive processes like Attention, Mental Recall, Orientation, and Learning, through Arousal to Fear and Adaptation. A smaller group associated with health workforce issues centred on Attitude of Health Personnel radiating out to Nurse-Patient Relations, Physician's Role, Consumer Behavior, and Job Satisfaction.

1987-91 A completely connected network driven by preparations for the DSM-IV published in 1994 continues the dominance of neuropsychological processes with a central axis of Attention and Arousal closely associated with Problem Solving, Orientation, and Cognition Disorders, and a more distant grouping associated with the diagnostic and epidemiological work underlying the DSM-IVs nosology including Psychiatric Status Rating Scales, Personality Tests, and Personality Development.

1992-96 One of the two peaks of emerging topic activity possibly associated with the DSM-IV, this period is dominated by an axis linking Patient Satisfaction and Health Knowledge, Attitudes, Practice, consistent with the growing importance of community based psychiatric care and the recovery movement. The appearance of Assisted Suicide and previously less investigated phenomena including Erectile Dysfunction and Male

Homosexuality indicate the growing breadth of the bio-psycho-social model first detectable in the mid-to-late 1970s.

1997-2001 Two reasonable sized clusters in this period indicate the growing sophistication and importance of cognitive models, with the most important keyword, Cues, repeated from the previous period but now representing both the largest group of articles, and forming the central part of the dominant network focused on explaining complex clinical conditions such as Cocaine-Related Disorders and Tobacco Use Disorder as the result of cognitive processes such as Psychological Inhibition, Judgment, and Social Identification/Self Efficacy. The smaller cluster is more focused on service characteristics including Nurse's Role, Personal Autonomy, and Disclosure.

2002-2006 Another of the unitary networks, with two semi-networks each focused on one of the two largest nodes. One side of the network, grouped around the largest keyword (Recognition, Psychology), appears to reflect another stage in the development of psychological theories including Psychophysics, Association Learning, Concept Formation, and Comprehension Tests/Language Tests. On the other side, the second and third largest keywords, Social Perception and Personality Inventory are linked with more therapeutic concepts used in psychotherapy including Trust, Empathy, Intention, and Narration.

2007-2011 The final unitary network is centred on another axis probably driven by the development of the DSM-V, with the two largest nodes being Neuropsychological Tests and Psychiatric Status Rating Scales. Arrayed around these nodes are smaller groupings including largely unconnected assessment-oriented features such as

	Psychological Interview, Executive Function, and Child Behavior, and a verbal process cluster comprising Language Tests, Psycholinguistics, and Speech.
2012-2016	The most recent network is unusual in that the keyword with the greatest number of articles in the period (Cognitive Dysfunction) is relatively unconnected to most of the other nodes, forming a small cluster with Spatial Memory. The rest of the nodes are closely linked around a set of concepts which show the growing importance of social factors in psychiatric/psychological theory and practice, with one end of an axis – Suicidal Ideation, linked with Bullying, Bisexuality, Psychological Resilience, and Mindfulness; connected via the other end of the axis – Social Stigma, with specific risk factors including Racism, and a substance abuse subnetwork including Opioid Related Disorders, Tobacco Use Disorder, Gambling, and Marijuana Smoking.

In the most recent period, attention to emerging topics in the published psychiatric/psychological literature would have alerted a curriculum development project to the growing importance of social phenomena, stigma, vulnerable populations including minorities, and complex determinants such as bullying and professional burnout, in the effort to address suicide and parasuicidal symptoms including suicidal ideation. In addition to the counterbalance to biases which may influence expert-driven curriculum development, an empirical approach like the current one has the advantage of being directly linked with the published literature. Using Web of Science, a topic search for “social stigma” and “suicidal ideation” sorted by times cited in the years 2012-2016 returns an article on mechanisms of risk for depression and suicidal ideation among LGB youth which could be used in the curriculum development process (Baams et al., 2015).

The techniques described here cannot replace expert judgement in curriculum development but can be used by experts involved in curriculum development to identify topics to be considered for inclusion. The model used to create MeSH by the NLM shows how the information provided by data mining techniques can be integrated into a system where human experts use the outputs of data mining techniques in combination with expert judgement to achieve better outcomes than would have been achieved by either expert or technique alone (Mork et al., 2013). Another useful model of how these techniques could be integrated into curriculum development is provided by Chen & Song (2019). These authors used co-word analysis to explore how systematic reviews can optimally leverage bibliometric techniques to avoid missing relevant research, and to compare the effectiveness of alternative search strategies in returning relevant and specific results. They describe the seminal work of Swanson, who is recognised as an early pioneer of literature-based discovery (LBD) for his work linking Raynaud's phenomenon to a potential treatment purely by a systematic approach to the examination of the relevant literature, rather than by empirical investigation (Swanson, 1993). Their approach would be particularly useful for the curriculum development process by making it very easy to trace the pathways of scientific development across years and citations.

5.4.2 - Complementary datasets

The information contained in the published literature is not the only large set of independent empirical evidence that would be useful for the curriculum development process with judicious bibliometric analysis. Although we have started with indexed literature due to its public availability and highly structured format, there is a wealth of other data that could be used to address biases. Complementary to the theoretical and experimental evidence contained in the literature, electronic medical records contain a wealth of detailed information about what is actually done in practice across all medical and allied health disciplines. Government and other public entities including medical licensing bodies,

pharmaceutical organisations, and health services record a great deal of data about diverse variables including prescribing practices, economic indicators, epidemiological factors, and patient behaviours like engagement, adherence, and changing demands. Morbidity and mortality registers track the most common and most serious causes of negative outcomes associated with health care. A curriculum informed by systematic mining of all available sources of information could reduce expert-driven biases and constantly calibrate educational experiences to match the real-world importance of the constantly changing features of medical practice.

5.4.3 - Complementary methods

Our research focused on a single method chosen for ease of interpretation and public access to data. It has been proposed that alternative bibliometric techniques such as co-citation, co-word, co-author, and bibliographic coupling, reveal different characteristics of scientific knowledge, and we are still in the early stages of determining which methods are most suited to which tasks (Boyack & Klavans, 2019). An ideal combination of techniques for the purposes of curriculum development might chain the technique reported in this paper with that reported by Chen & Song (2019). Such a combination could use the Increment rate statistic to identify emerging topics, and Chen & Song's cascading citation expansion to outline the development of strands of scientific knowledge from seminal articles to the most important recent publications (Chen & Song, 2019). Li et al (2022) report a co-word analysis that demonstrates how these techniques can be used to understand the thematic evolution of the concept of psychological distance across four distinct periods, from infancy, through exploration and growth, to the outbreak to more general dissemination (Li et al., 2022).

Use of the MeSH controlled vocabulary rather than co-word analysis of free text has the potential disadvantage that there is likely to be a delay between the reporting of new ideas, methods, or conclusions in the literature, and their incorporation into the vocabulary. Co-word analysis of the free-text of published articles' title, abstract, or body, may be the method with the potential for the earliest identification of emerging topics, although as noted the polysemous nature of free-text can make this type of analysis difficult to interpret.

5.4.4 - Limitations and Future Research

In addition to issues noted in the discussion above, the research reported here displays the limitations of exploratory research. Identifying emerging topics by the Increment statistic is a first step towards a framework that uses objective evidence to reduce biases in curriculum development. It is encouraging that the emerging topics identified are consistent with our knowledge of the psychiatric and psychological domains and appear to show the anticipated effect of the publication of different editions of the DSM. However, confidence in the validity of the information would be improved by empirical evidence that it can be effectively integrated with expert judgement. We are developing an experimental paradigm that will present the emerging topics to a group of psychiatrists engaged in continuous professional development to confirm the novelty of the knowledge identified, and to calibrate the most useful form for the presentation of the information for use in curriculum development. Further research would be required to estimate or measure the impact of this information on biases exhibited by experts during curriculum development and related activities.

The other main limitations of the paper are related to choices regarding methods and materials. As discussed in the text, we have chosen the Increment statistic over other potential methods for identifying emerging topics such as citation analysis; the MEDLINE database over larger databases such

as Scopus or Web of Science; and the MeSH controlled vocabulary rather than free text. These choices have the advantages of a database focused on relevant clinical research; open access to data; and the ability to identify emerging topics without having to wait for citation paths to become evident. As there are likely to be countervailing advantages to other materials and methods, if it can be demonstrated that integrating evidence about emerging topics in curriculum development processes can reduce expert bias, it may prove useful to compare the choices we have made with the alternatives. For example, the coverage of a larger set of literature provided by Web of Science or Scopus might prove better at identifying and addressing gaps in the coverage of medical curricula.

5.5 - Conclusions

We have argued that data mining techniques used in bibliometrics may be valuable for reducing biases implicit in expert-driven medical curriculum development by identifying the most important emerging topics independent of expert biases. We have shown that a network analysis based on the Increment rate statistic computed from MeSH annotations of the medical literature can identify the most active emerging topics in psychiatry and psychology, suggest which components of a psychiatric curriculum and syllabus to review, and identify specific articles upon which to base the review. This is the only current alternative to expert judgement for the identification of topics to be considered during curriculum development. While the published literature is a natural first target for the application of data mining techniques to curriculum development due to its large, structured, and freely available data set, other less accessible sources would provide objective data about the relative importance of a wide range of the knowledge, skills, and attitudes needed for an up-to-date curriculum.

5.6 - References

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). American Psychiatric Association.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorder* (4th ed). American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed). American Psychiatric Association.

Amos, A. J., Lee, K., sen Gupta, T., & Malau-Aduli, B. S. (2021). Systematic review of specialist selection methods with implications for diversity in the medical workforce. *BMC Medical Education*, *21*(1). https://doi.org/10.1186/s12909-021-02685-w

Åström, F. (2007). Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990-2004. *Journal of the American Society for Information Science and Technology*, *58*(7), 947–957. https://doi.org/10.1002/asi.20567

Baams, L., Grossman, A. H., & Russell, S. T. (2015). Minority stress and mechanisms of risk for depression and suicidal ideation among lesbian, gay, and bisexual youth. *Developmental Psychology*, *51*(5), 688–696. https://doi.org/10.1037/a0038994

Benson, N. M., Vestal, H. S., Puckett, J. A., Taylor, J. B., Hogan, C., Smith, F. A., & Beach, S. R. (2019). Continuous Quality Improvement for Psychiatry Residency Didactic Curricula. *Academic Psychiatry*, *43*(1), 110–113. https://doi.org/10.1007/s40596-018-0908-4

Boyack, K. W., & Klavans, R. (2019). Creation and Analysis of Large-Scale Bibliometric Networks. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 187–212). Springer Nature. https://doi.org/10.1007/978-3-030-02511-3_8

Braun, L., & Saunders, B. (2017). AMA Journal of Ethics ® PEER-REVIEWED CME ARTICLE: MEDICAL EDUCATION Avoiding Racial Essentialism in Medical Science Curricula (Vol. 19, Issue 6). www.amajournalofethics.org518

Brunn, M., Diefenbacher, A., Courtet, P., & Genieys, W. (2020). The Future is Knocking: How Artificial Intelligence Will Fundamentally Change Psychiatry. *Academic Psychiatry*, *44*(4), 461–466. https://doi.org/10.1007/s40596-020-01243-8

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, *22*(2), 191–235. https://doi.org/10.1177/053901883022002003

Chen, C. (2003). Visualizing scientific paradigms: An introduction. *Journal of the American Society for Information Science and Technology*, *54*(5), 392–393. https://doi.org/10.1002/asi.10224

Chen, C. (2006). CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. https://doi.org/10.1002/asi.20317

Chen, C., & Song, M. (2019). Visualizing a field of research: A methodology of systematic scientometric reviews. *PLoS ONE*, *14*(10). https://doi.org/10.1371/journal.pone.0223994

de Langhe, R. (2017). Towards the discovery of scientific revolutions in scientometric data. *Scientometrics*, *110*(1), 505–519. https://doi.org/10.1007/s11192-016-2108-x

D'Eon, M., & Crawford, R. (2005). The elusive content of the medical-school curriculum: A method to the madness. *Medical Teacher*, *27*(8), 699–703. https://doi.org/10.1080/01421590500237598

Harden, R. M. (2001). AMEE Guide No. 21: Curriculum mapping: A tool for transparent and authentic teaching and learning. *Medical Teacher*, *23*(2), 123–137. https://doi.org/10.1080/01421590120036547

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Kulasegaram, K., Mylopoulos, M., Tonin, P., Bernstein, S., Bryden, P., Law, M., Lazor, J., Pittini, R., Sockalingam, S., Tait, G. R., & Houston, P. (2018). The alignment imperative in curriculum renewal. *Medical Teacher*, *40*(5), 443–448. https://doi.org/10.1080/0142159X.2018.1435858

Li, S., Chen, H., Feng, Y., Chen, F., & Hou, C. (2022). Research Progress and Thematic Evolution of Psychological Distance—A Co-Word Analysis Based on Bibliometric Research. *Current Psychology*, *41*(3), 1569–1583. https://doi.org/10.1007/s12144-020-00690-8

Moral-Munoz, J. A., López-Herrera, A. G., Herrera-Viedma, E., & Cobo, M. J. (2019). Science Mapping Analysis Software Tools: A Review. In *Springer Handbook of Science and Technology Indicators* (pp. 159–185). Springer Nature. https://doi.org/10.1007/978-3-030-02511-3_7

Mork, J. G., Yepes, A. J. J., & Aronson, A. R. (2013). *The NLM medical text indexer system for indexing biomedical literature*. CEUR Workshop Proceedings. https://ii.nlm.nih.gov/Publications/Papers/MTI_System_Description_Expanded_2013_Accessible.pdf

Ohniwa, R. L., & Hibino, A. (2019). Generating process of emerging topics in the life sciences. *Scientometrics*, *121*(3), 1549–1561. https://doi.org/10.1007/s11192-019-03248-z

Ohniwa, R. L., Hibino, A., & Takeyasu, K. (2010). Trends in research foci in life science fields over the last 30 years monitored by emerging topics. *Scientometrics*, *85*(1), 111–127. https://doi.org/10.1007/s11192-010-0252-2

Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science*, *45*(1), 31–38. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<31::AID-ASI4>3.0.CO;2-G

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org/index.html

Risberg, G., Johansson, E. E., & Hamberg, K. (2009). A theoretical model for analysing gender bias in medicine. *International Journal for Equity in Health*, *8*. https://doi.org/10.1186/1475-9276-8-28

Roberts, C., Khanna, P., Rigby, L., Bartle, E., Llewellyn, A., Gustavs, J., Newton, L., Newcombe, J. P., Davies, M., Thistlethwaite, J., & Lynam, J. (2018). Utility of selection methods for specialist medical training: A BEME (best evidence medical education) systematic review: BEME guide no. 45. *Medical Teacher*, *40*(1), 3–19. https://doi.org/10.1080/0142159X.2017.1367375

Swanson, D. R. (1993). Intervening in the life cycles of scientific knowledge. *Library Trends*, *41*(4).

Swanwick, T., Forrest, K., & O'Brien, B. C. (Eds.). (2018). *Understanding medical education: evidence, theory, and practice* (Third). Wiley Blackwell. https://doi.org/10.1002/9781119373780

The Lancet. (2019). Cardiology's problem women. *The Lancet*, *393*(10175), 959. https://doi.org/10.1016/S0140-6736(19)30510-0

Thomas, P. A., Kern, D. E., Hughes, M. T., & Chen, B. Y. (Eds.). (2015). *Curriculum development for medical Education a six-step approach* (3rd ed.). Springer Publishing Company, Inc.

Thomas, P., Kern, D. E., Hughes, M. T., & Chen, B. Y. (Eds.). (2015). *Curriculum Development for Medical Education: A Six-Step Approach* (Third). Johns Hopkins University Press.

van Raan, A. (2019). Measuring Science: Basic Principles and Application of Advanced Bibliometrics. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 237–280). Springer Nature. https://doi.org/10.1007/978-3-030-02511-3_10

Vernon, M. M., Danley, C. M., & Yang, F. M. (2021). Developing a measure of innovation from research in higher education data. *Scientometrics*, *126*(5), 3919–3928. https://doi.org/10.1007/s11192-021-03916-z

Wang, Y., Zhao, Y., Zheng, J., Zhang, A., & Dong, H. (2018). The evolution of publication hotspots in the field of telemedicine from 1962 to 2015 and differences among six countries. *Journal of Telemedicine and Telecare*, *24*(3), 238–253. https://doi.org/10.1177/1357633X17693749

World Psychiatric Association. (2002). World Psychiatric Association Institutional Program on the Core Training Curriculum.

Zitt, M., Lelu, A., Cadot, M., & Cabanac, G. (2019). Bibliometric Delineation of Scientific Fields. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer Handbook of Science and Technology Indicators* (pp. 25–68). Springer Nature. https://doi.org/10.1007/978-3-030-02511-3_2

Chapter 6: Novel sparse matrix algorithm expands the feasible size of a self-organizing map of the knowledge indexed by a database of peer-reviewed medical literature

Authors:

Andrew James Amos[*1], MB.BS, Kyungmi Lee[2], PhD, Tarun Sen Gupta[1], PhD, Bunmi S. Malau-Aduli[1,3], PhD

* Corresponding Author

1 College of Medicine and Dentistry, James Cook University, Townsville, Australia

2 College of Science and Engineering, James Cook University, Cairns, Australia

3 School of Medicine and Public Health, University of Newcastle, Newcastle, Australia

This chapter describes the innovations to the sparse form of the self-organising map algorithm that allowed for the integration of the entire Medline database of ~30 million articles and ~29 thousand medical subject headings (MeSH) as the training corpus in place of the smaller subsets used by previous research.

6.1 - Abstract

Past efforts to map the Medline database have been limited to small subsets of the available data because of the exponentially increasing memory and processing demands of existing algorithms. We designed a novel algorithm for sparse matrix multiplication that allowed us to apply a self-organizing map to the entire Medline dataset, allowing for a more complete map of existing medical knowledge. The algorithm also increases the feasibility of refining the self-organizing map to account for changes in the dataset over time.

6.2 - Introduction

The quantity of information available to modern researchers for analysis, both structured data with defined meaning and unstructured information of uncertain meaning, is increasing exponentially.[1] However, it is impossible for any individual human being to read and understand more than a tiny fraction of the information that exists at a point in time, or to keep up with new information as it becomes available.

Fortunately, a wide variety of machine learning (ML) algorithms have been developed that can analyze and present abstract summaries of characteristics of arbitrarily large datasets relevant for particular purposes. Visualization is one of the most powerful ML techniques for condensing the information contained in large datasets into human understandable form. The most common approach is to transform high dimensional data into a flat two dimensional map which retains the properties of interest in the high dimensional space.[2]

An important example of the visualization of high dimensional data in two dimensional maps is the production of maps of scientific knowledge produced in the Science of Science field (SciSci).[3] SciSci maps have been created to summarize a wide range of relationships including research networks linking

scientists, scientific units such as research labs, universities, and more abstract information such as the emergence and evolution of new scientific ideas.[4]

The most common form of SciSci visualization is citation analysis which represents abstract relationships between scientific articles published in peer-reviewed journals.[4] This type of analysis is particularly useful for identifying the evolution of scientific paradigms, comparing the relative productivity of different researchers or research units such as universities, and analyzing the impact of network dynamics such as multidisciplinarity on research productivity measured by quantity and quality.[5]

6.3 - Visualization of large medical article datasets with self-organizing maps

The most comprehensive and authoritative source of information about medical knowledge and practice is contained within the articles published by peer-reviewed medical journals. While there are other stores of medical knowledge such as textbooks, manuals, and online learning materials, those types of knowledge rely upon references to peer-reviewed articles as authority for their claims.

Unlike other forms of medical knowledge, articles published in peer-reviewed journals have a systematic structure and quality-control mechanisms. In addition, they are labelled and organized to facilitate rapid identification and retrieval of articles relevant to particular information needs. All these features make the set of peer-reviewed medical articles an excellent test case for developing SciSci visualization techniques.

Skupin has provided the most advanced visualizations of the peer-reviewed medical literature. His maps combine the powerful pattern recognition of self-organizing maps (SOMs) with the sophisticated visual cues of map-making to produce high density visual abstractions of the domains of medical knowledge (Figure 6-1).[6]

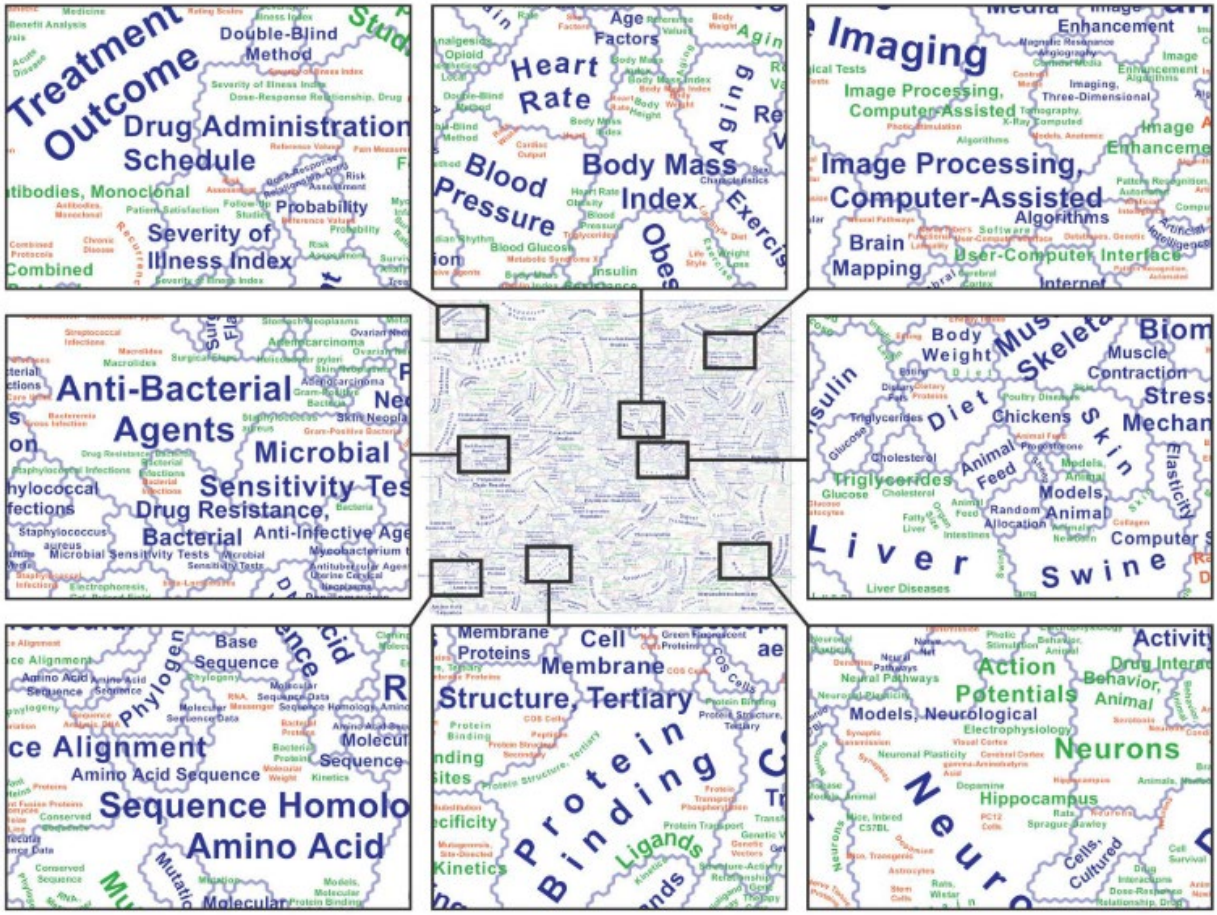


Figure 6-1: Skupin's map of the medical literature showing regions of interest is licensed under CC-BY (https://creativecommons.org/licenses/by/4.0/)[6]

Due to hardware and software limitations, Skupin's SOM was trained on 2.1 million of the more than 20 million indexed peer-reviewed articles available at the time. In addition, the size of the SOM itself was limited to a two-dimensional grid of neurons of size 275 x 275 for a total of 75,625 neurons. Skupin conceded that it would have been preferable to increase the dimensions of the grid to have a smaller article to neuron ratio and improve the resolution of the map, but that this was not computationally feasible.[6]

Skupin et al.[6] acknowledged that the size of their map limited its ability to reproduce detailed structures of the high-dimensional knowledge domain on the two-dimensional surface of the SOM. In addition to the limitation of being forced to train the SOM on a sample of about 10% of articles rather than the complete set, they were only able to consider a small subset of the medical subject headings (MeSH – described below) used to annotate each article. As a result, the Skupin SOM is unable to represent any information that was contained in the 90% of articles and MeSH not used for training.

Another limitation of a small SOM and partial training set is that it constrains the complexity of the relationships that can be represented. Figure 6-1 illustrates this point.[6] Each of the 8 rectangles surrounding the central map represents a cluster of related knowledge structures. For example, the bottom right rectangle represents the high-level specialty Neurology, superimposed on lower level anatomical structures like neurons and hippocampi, as well as neural processes like action potentials.

The exclusion of most of the MeSH from training means that their map cannot contain most of the categories that annotate Medline-indexed articles. For example, the Skupin map might contain the high-level MeSH structure Hippocampus, but not the lower-level MeSH structure CA1 Region, Hippocampus; or it might contain the high level MeSH-labelled process Action Potentials, but not the related process of Synaptic Potentials. As the MeSH controlled vocabulary has a hierarchical structure with high level concepts branching into multiple lower-level concepts, this necessarily means that the Skupin SOM is limited to representing relationships between relatively high level MeSH.

A final limitation imposed by computational feasibility is the inability to extend the SOM to consider more than two dimensions. While it is technically possible to add a third spatial dimension to a SOM, it requires special equipment to make use of the results.[7] However, the ability to consider how a two dimensional spatial SOM changes over time promises to reveal the dynamic processes involved in the emergence and evolution of knowledge structures.[8] Any technique that increases the computational

efficiency of self-organising map training also improves the possibility of extending the regular model to reveal new processes such as temporal relationships.

6.3.1 - Visualization with self-organizing maps

The self-organizing map pioneered by Kohonen has been used to visualize subsets of the Medline database which indexes and annotates more than 30 million articles from high-quality peer reviewed medical journals. Kohonen's iterative algorithm has many attractive properties for generating visualizations of large sets of scientific knowledge. It can effectively map very high dimensional information spaces to two dimensions while retaining many of the topological relationships of the higher dimensional space. In addition, it can detect and help visualize previously unknown relationships.[2]

Self-organizing maps are single-layer neural networks which transform high-dimensional vectors into low-dimensional vectors (usually 2D) while retaining many of the topological properties of the higher dimensional space (Figure 6-2).[2,9] The most obvious example of the retention of topological properties by SOMs is that units of information that are close together in high-dimensional space will also be close together in the 2D space represented by the SOM.

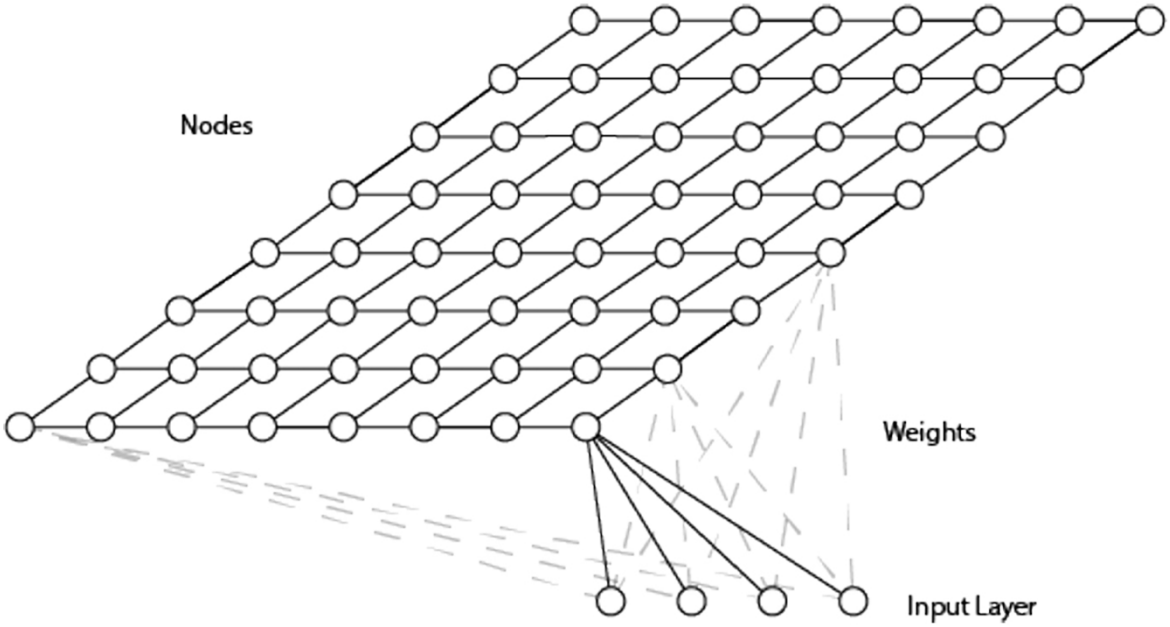


Figure 6-2: Self-organizing map with a 4-dimensional input layer and a 2-dimensional output layer of 64 nodes

This retention of topological properties by SOMs is extremely useful for visualization because human beings have visual systems that are highly developed for discriminating spatial relationships in 2D space but have almost no ability to perceive similar relationships in high-dimensional space. As a result, SOMs can make complex relationships from high dimensional spaces intelligible to human vision.

Kohonen[2] described two algorithms for training SOMs to transform high-dimensional input vectors to 2D neuronal maps. Each neuron in the SOM is represented by a vector of weights of the same dimensionality as the input vectors. Before training the weight vectors of all neurons are filled with random values. Training occurs in distinct epochs during which every input is separately presented to the entire network of neurons and the distance between each input vector and each neuron is calculated by a standard distance function, often the squared Euclidean distance. Each input is linked

with the neuron that is closest to it, and training is implemented by iterative changes that move neuronal weight vectors towards the inputs that they are closest to.

In the standard SOM algorithm training changes are made to weights after every individual input is presented, while in the batch SOM algorithm changes are accumulated over the entire epoch and applied at the end of the epoch. The batch algorithm is computationally more efficient than the standard algorithm, and is thought to converge on the same set of training outcomes, although this has not been formally proved.[2]

6.3.2 - Improving computational efficiency for sparse input vectors

Other techniques are used to improve the computational efficiency of SOM algorithms for specific types of input. A technique that is useful for mapping the medical literature is a form of the batch algorithm that is optimised for sparse inputs. An input vector can be considered sparse when most of the input elements are empty (or where most of the elements have a single constant value such as 0 or 1 that can effectively be treated as empty).[10,11]

Sparse inputs present a challenge for GPU-based computation because they complicate the data-partitioning that is required to distribute tasks across the large number of processing units. Melka & Mariage[10,11] described a modified version of the batch SOM algorithm optimised for sparse inputs and suitable for GPU processors (Figure 6-3). Essentially their algorithm depends on the fact that only non-zero input elements affect the weights of each neuron in the SOM which allows them to precompute the values of the squared norms of the distance equation once for the input vectors, and once per epoch for the weights of each neuron, substantially reducing the computational load.

Variables/Data structures

	Input: x a set of *N* sparse vectors of *D* components
	Data: ω initialized codebook of M dense vectors
	Data: χ an array of *N* reals, satisfying $\sum_j x_{ij}^2$
	Data: *dst* array of N reals to store best distances
	Data: *bmu* array of N integers to store best match units
	Data: *nmu* array of D reals to accumulate numerator values

Algorithm

1	**for** i ← 1 **to** N **do** $\chi_i \leftarrow \sum_j x_{ij}^2$; init χ
2	**For** *e* ← 1 **to** e_{max} **do** train one epoch
3	interpolate *σ*;
4	**for** *i* ← 1 **to** *N* **do** $dst_i \leftarrow \infty$; initialize *dst*
5	**for** *k* ← **to** *M* **do** find all bmus
6	$\omega \leftarrow \sum_j x_{kj}^2$
7	**forall** *i* ∈ 1,...,*N* do
8	d ← $\omega + \chi_i - 2(x_i \cdot \omega)$;
9	**if** *d* < dst_i **then** store best match unit
10	$dst_i \leftarrow d$;
11	$bmu_i \leftarrow$ k;
12	**forall** *k* ∈ 1,...,*M* do
13	*den* ← 0 ; init denominator
14	**for** *j* ← 1 **to** *D* **do** $num_j \leftarrow 0$; init numerator
15	**for** *i* ← 1 **to** *N* **do** accumulate *num* and *den*
16	c ← bmu_i ;

17	h ← exp($\frac{\|r_k - r_e\|^2}{2\sigma^2}$
18	*den* ← *den* + *h* ;
19	**for** *j* ← 1 **to** *D* **do**
20	$num_j \leftarrow num_j + h x_{ij}$
21	**for** *j* ← 1 **to** *D* **do** update ω_k
22	$\omega_{kj} \leftarrow \frac{num_j}{den}$

Figure 6-3: Sparse batch SOM algorithm[10,11]

6.3.3 - Optimising the SOM sparse algorithm for nominal inputs

6.3.3.1 - The Medline database

The SOM algorithm described by Melka & Mariage[10,11] is designed for sparse inputs comprising ratio variables (i.e. real numbers with arbitrary limits). Some data sets involve sparse inputs with nominal variables which allow for further optimization of the batch algorithm that increase computational efficiency and decrease the memory requirements for storing and distributing the input vectors.

The Medline database used by Skupin to create SOMs of the knowledge contained within the peer-reviewed medical literature is an example of a database with nominal input vectors.[6] The Medline database indexes a detailed set of information for each article published in a selected group of high-quality peer-review medical journals. It includes articles all the way back to the 19th century, but coverage is limited prior to the 1970s.[12]

For each indexed article Medline records the title, abstract, year of publication, and authors, alongside many other variables.[13] Most relevant to the SOMs produced by Skupin, Medline annotates every published article with a set of descriptors called Medical Subject Headings (MeSH) from a controlled

vocabulary maintained by the National Library of Medicine. MeSH describe the content, methods, and other characteristics of each article (such as research design, patient population, and disease treated).

For example, an article describing a treatment trial where patients with schizophrenia were treated with the antipsychotic medication risperidone and compared with controls treated with placebo would be annotated with MeSH including the diagnosis of Schizophrenia, the experimental design Randomised Control Trial, and the treatment group Antipsychotic – Risperidone.

Skupin's SOM used the MeSH annotating each article in the Medline database as an input vector, where the nominal MeSH variables were represented as a vector of binary elements.[6] Skupin selected the most frequently used 10% of the 23,000 MeSH in the controlled vocabulary for the inputs to his SOM. Each article in the Medline database was then characterised by a vector with 2,300 elements that were coded as either present or absent (numerically as 0 or 1). As articles in the Medline database are annotated with an average of 10 MeSH, the input vectors were highly sparse, with around 2,290 elements indicating a MeSH was absent and around 10 elements indicating a MeSH was present.

Compared with the real variable inputs to Melka & Mariage's batched sparse SOM,[10,11] the binary inputs of the Medline dataset represent an additional opportunity for optimisation. While various structures are used for compactly storing sparse matrices and algebraically combining them with other sparse and dense matrices, they generally require at least two vectors – one to describe the position of a non-zero element, and a second to describe its value. The binary input vectors representing Medline articles can dispense with the second vector.

In addition, because all elements in the sparse Medline input vectors are either zero or one, the dot-product used to calculate distances in the SOM algorithm can be replaced by simple addition. Combining these two modifications, Amos et al.[14] described an algorithm (Figure 6-4; the Batched sparse binary SOM - MedSOM) that increased the computational efficiency and data compression sufficiently to

create a SOM trained on the entire set of more than 33 million articles indexed in Medline as of 01/01/2021; and included all 29,917 of the MeSH in the NLM's controlled vocabulary.

Variables/data structures	
	Input: x: *N* sparse vectors of *D* elements representing Medline articles
	Data: ω: initialized codebook of *M* dense vectors of length *D*; represents weights between inputs and each SOM node
	Data: χ: array of *N* reals; as x is sparse this matrix is the number of non-zero elements per row (i.e., the number of MeSH annotating each article)
	Data: *dst*: array of *N* reals storing distance to best matching unit for each article
	Data: bmu_1, bmu_2 – 2 vectors of *N* integers to store best matching and second-best matching unit/node for each article
	Data: *num* – array of *D* reals to accumulate numerator values

Algorithm	
1	**Randomize** codebook weights ω between 0.0f and 1.0f
2	**for** each epoch *e* ← 1 to *K* **do**
3	**compute** σ; // radius for current epoch
4	**standardize** codebook weights; sqrt of sum of weights per node
5	**calculate** bmu_1, bmu_2 for each article *i* ← 1 to *N*
6	**for** each node *m* ← 1 to *M*
7	**sum** $m \cdot i$ // dot-product replaced by sum as all $i = 1$
8	**calculate** $dst = D_a - 2 \cdot (m \cdot i)$
9	**calculate** new weights for each node ← 1 to *M* **do**
10	**for** each article *i* ← 1 to *N*

11	**calculate** $h(e) = \exp\left(-\frac{\|r_k - r_c\|^2}{2\sigma(e)^2}\right)$ // Neighborhood function
12	**reduce** denominator = accumulate h_{ck} per node
13	**calculate** numerators = accumulate h_{ck} per node/article
14	**calculate** new weights = numerators / denominator
15	**for** $i \leftarrow 1$ to N
16	**calculate** adjacency = ($bmu_1 - bmu_2 =< 1$) // Manhattan distance
17	**calculate** topographic error = % adjacent

D – total number of MeSH categories (=29,917); D_a – number of MeSH categories annotating an individual article; M – number of nodes (350 x 350 = 122,500); N – number of articles (=33,375,866); *h(e)* – width of the neighborhood (which changes over training epochs according to the formula $175/(1.7)^{epoch}$)

Figure 6-4: Sparse batch SOM algorithm for binary/nominal input vectors (MedSOM)

Starting with the 275 x 275 = 75,625 neuron SOM used by Skupin to model the subset of 2.1 million articles with 2,300 MeSH, Amos et al. experimented with different sizes of SOM in increments of 25 x 25 up to 400 x 400 = 160,000 neurons and settled on a 350 x 350 = 122,500 neuron SOM. Maps larger than this did not achieve any greater reduction in topographic error, which measures the % of articles where the best matching unit and second-best matching unit/node are not adjacent.[15]

6.3.3.2 - Iterative improvement of the SOM algorithm

The LibSVM and MedSOM algorithms seek to improve the efficiency of the SOM algorithm by using the properties of sparse matrix multiplication. In addition, by reducing the space required to represent the

training articles in GPU memory, the MedSOM algorithm makes it possible to leverage GPU specific optimisations of matrix multiplication.

The current implementation of all the algorithms used in this research use general memory accessible to all GPU threads rather than the cache memory accessible only within thread blocks. This substantially slows processing speed because it requires atomic calls to article information in general memory for each MeSH representing each article.

In general, the speed of implementation of GPU algorithms in programming kernels is considered to be either compute bound, where the time taken to complete each instruction set depends mainly on the time taken to complete the computational instructions; or memory bound, where the time taken to complete each instruction set depends mainly on the time taken to copy the data required to complete the instructions from other parts of memory to the thread block where it is required.[16]

Reducing the size of memory required to represent articles by moving from dense to sparse representation of articles makes it feasible to optimise processing by copying the representation of articles into each thread block, eliminating the need for atomic operations that pause computation while retrieving individual items from general memory. Reducing the memory requirement further by moving from the LibSVM to the MedSOM algorithm increases the number of articles that can be copied into the local cache of each thread block. Both reduce the possibility/extent of being memory bound and compute bound for any given set of Medline data.

6.3.3.3 - Capitalizing on improved performance with model extensions

In addition to making larger and more complex SOMs possible, increasing the speed and decreasing the memory requirements of the SOM make it more feasible to create extensions that expand the capacities

of the basic model. One of the main limitations of the existing MedSOM is that it was trained with articles across more than a century of research. While it is likely there were some features of medical practice in 1900 that remain true today, it is certain that there are many features that have changed radically. The MedSOM currently has no way of differentiating between such features.

One extension of the SOM that would allow MedSOM to detect changes in the peer-reviewed medical literature over time is Denny et al.'s relative density SOM (RedSOM).[8] By training a SOM on each year of their dataset and then analyzing differences in the map between years, Denny et al. were able to visually identify "emerging clusters, disappearing clusters, split clusters, merged clusters, enlarging clusters, contracting clusters, the shifting of cluster centroids, and changes in cluster density" (p281).[8] For example, by this technique it would be possible to identify the emergence of clusters of new research activity, and the disappearance of clusters of old research activity.

The goal of Amos et al.'s[14] MedSOM was to provide a tool capable of providing an empirical basis for curriculum development, in order to address the biases inherent in the current reliance on expert judgement. While MedSOM's knowledge map of the entire set of information contained in more than a century of peer reviewed medical literature was the first step towards this goal, the ability to differentiate between emerging and obsolete information is clearly vital for informing a curriculum.

6.4 - Performance evaluation

The optimised algorithm described by Amos et al.[14] is important for two reasons. By decreasing the size of the sparse input vector, it increases the size of the dataset that can be physically held in memory and used to train the SOM describing the published medical knowledge. By simplifying the calculation of the distance between each article and each neuron in the SOM it increases the size of the SOM that can be trained in realistic time by current technology.

We set out to demonstrate both these factors by comparing the performance of the Amos et al.[14] MedSOM algorithm to Melka & Mariage's[10,11] BSOM algorithm by calculating the maximum size of the input/weight vectors, SOM, and article input set possible with each. We then compared the speed performance of each algorithm on an epoch of training using Amos et al. as the baseline. We focused on the matrix multiplication required to calculate the best matching unit for each article, which was the most computation- and time-intensive part of the algorithm.

6.4.1 - Technical specifications

The analyses were performed on a Windows 11 PC with an Intel(R) Core(TM) i7-14700F 2.10 GHz CPU, 32GB RAM, and an NVIDIA GeForce RTX 4090 with 24GB RAM GPU.

6.4.2 - Memory

To analyse the influence of the sparse Melka and MedSOM algorithms on the feasible size of the SOM compared with a dense SOM algorithm we created a series of self-organising maps and sets of articles represented either by dense matrices of 4-byte floats with most elements set to 0.0 and annotated MeSH set to 1.0 (Dense-SOM); or by sparse matrices in either the Lib-SVM format which includes one 4-byte integer that identified each present MeSH, and a 4-byte float that represented the value of the MeSH; or the MedSOM format that included one 4-byte integer that identified each present MeSH. Each article was annotated with a random number of MeSH continuously distributed between 5 and 15.

Within this framework we compared the trade-offs and limits between the number of articles, number of MeSH, and SOM dimensions, across the dense and sparse algorithms. The results are presented in Figure 6-5.

When interpreting Figure 6-5 it is important to note that 5(a) is different from 5(b, c) in two ways. In order to illustrate the differences between the memory usage of the sparse algorithms, 5(a) represents the full range of 30 million articles. As can be seen, this flattens the memory usage of the Dense algorithm against the upper left limits of the graph, making it impossible to understand the trade-offs between MeSH number and article number; and it makes it difficult to see the memory usage of the Sparse algorithms for Articles, barely perceptible as a thin line at the bottom of the left limits of the graph.

In order to see the trade-offs between MeSH and articles for the Dense algorithm, 5(b) limits the range of articles between 0 and 1,200,000. As 3-dimensional graphs are easier to read when the figures are concave, the axes representing MeSH and Articles are switched. In order to see the differences in memory usage for the storage of articles by the Sparse algorithms, 5(c) limits the range of articles between 0 and 1,200,000, limits the range of memory usage from 0 to 110mb, and switches the MeSH and Article axes.

With those modifications in mind, 5(a) shows that both sparse algorithms enormously increase the capacity of self-organizing maps to represent both an increased number of Articles and an increased number of MeSH. For a SOM of dimensions 350x350, on the research equipment with 24GB of memory available, the dense algorithm was able to represent 1,000,000 articles annotated with 5,000 MeSH; or to represent 70,000 articles with 30,000 MeSH.

Figure 6-5(b) shows the trade-offs for the dense algorithm between number of MeSH and number of Articles, with the memory used to store the SOM increasing and memory used to store Articles decreasing as the dimensions of the SOM increased from 250x250 to 350x350.

Figure 6-5(a) also shows that the MedSOM algorithm significantly decreases the memory used to store Articles compared with the LibSVM algorithm. For a SOM of 350x350 nodes, representing 30,000 MeSH,

and trained on 30,000,000 articles, the MedSOM algorithm decreased the memory used from 3372mb to 1378mb relative to the LibSVM algorithm, with proportional decreases across the range of MeSH, articles, and SOM dimensions, a 59% reduction. Both sparse algorithms used the same memory to store the SOM across all dimensions.

Figure 6-5(c) shows that this advantage for MedSOM extended across the entire range of SOM dimensions, number of MeSH, and number of articles, with a 50% reduction from 12mb to 6mb to store articles with a SOM of 250x250 nodes, 30,000 MeSH, and 120,000 articles.

6.5 - Processing time

To analyse the influence of the sparse Melka and MedSOM algorithms on processing time we compared the time required to complete a single cycle of the matrix multiplication used to calculate the best-matching unit for each article ($(x_i \cdot \omega)$ on line 8 of Figure 6-3 and *(m · i)* on line 7 of Figure 6-4). The results are reported in Figures 6-6 & 6-7.

Figure 6-6 shows the massive advantage of the sparse algorithms over the dense algorithms in terms of processing speed. For a SOM of 350x350 nodes, memory constraints prevent it from handling more than 450,000 articles with 10,000 MeSH, or 70,000 articles with 30,000 MeSH. While the dense algorithm took 178 minutes to process one cycle of the best matching unit dot-product with these parameters, MedSOM took less than 1 minute. The same trade-offs between MeSH and article numbers seen for memory are replicated for processing speed, shown by the reduction in processing time per cycle to 90 minute gained by changing the article number to 70,000 and the MeSH to 30,000.

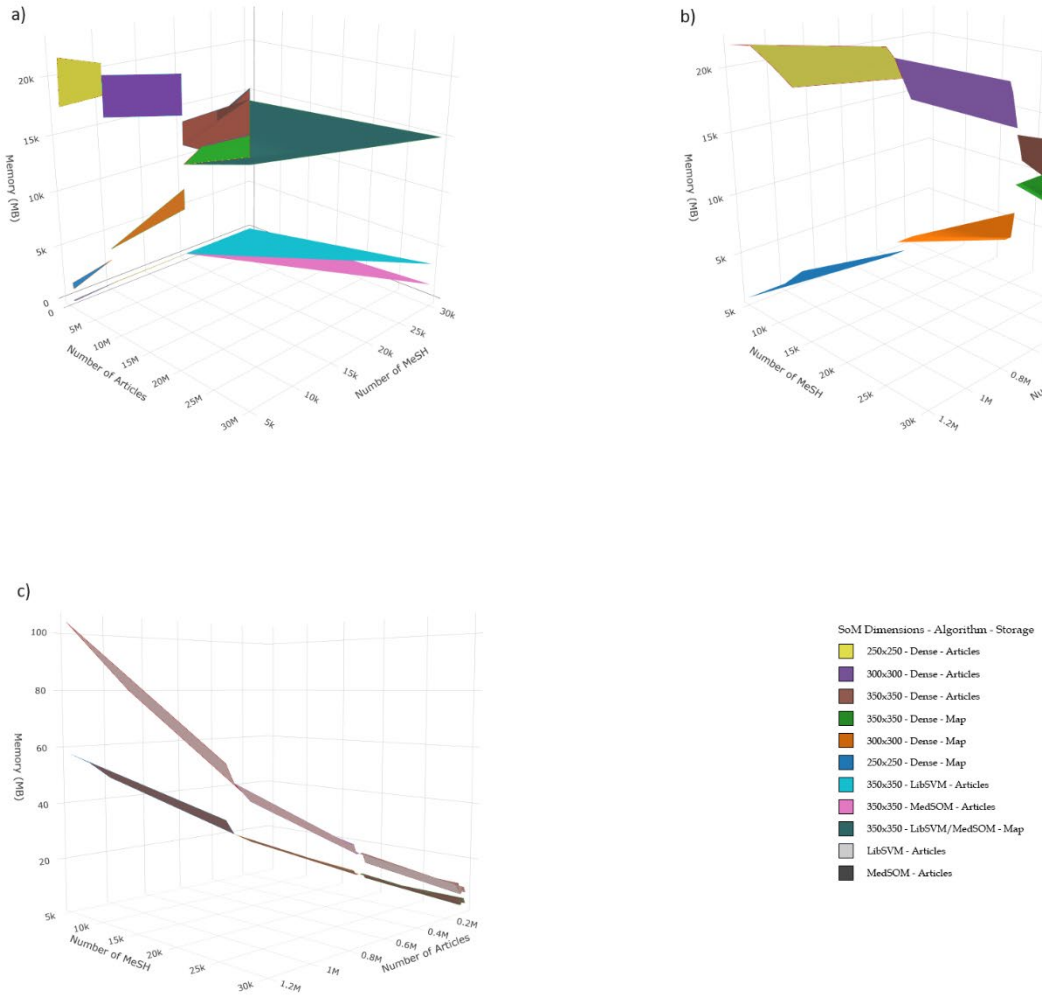


Figure 6-5 – Memory storage required for sparse and dense algorithms across conditions: a) Compares MB required to store articles and SOM for dense and sparse algorithms; b) Shows MB required to store articles and SOM for dense algorithm alone; c) Compares MB required to store articles for sparse algorithms

At the lower end, while the dense algorithm took 9 minutes to process a 350x350 node SOM with 50,000 articles and 5,000 MeSH, MedSOM took less than 1 second (554 times faster). Note that

differences between the MedSOM and LibSVM algorithms are not visible at the scale shown by Figure 6-6, so MedSOM is used to represent both.

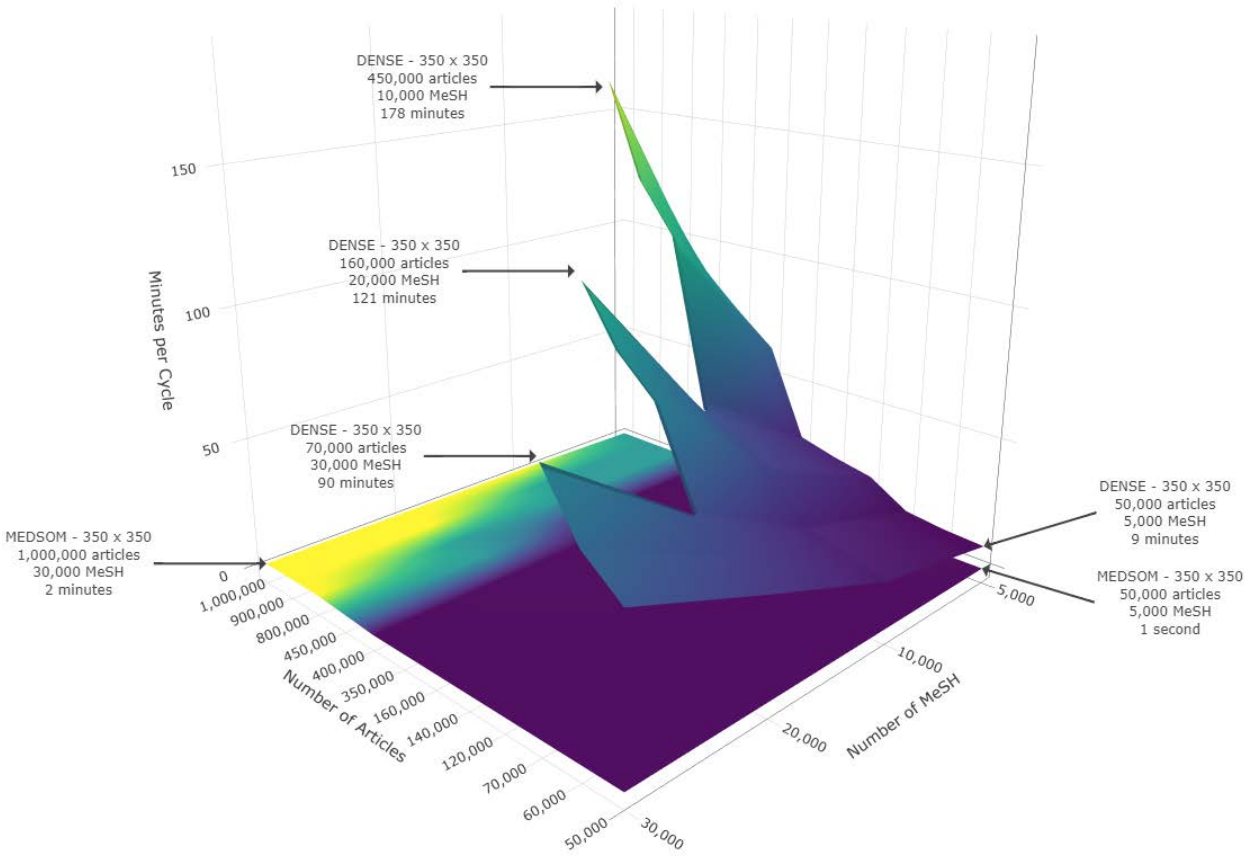


Figure 6-6: Processing time in minutes per cycle for the dense and MedSOM algorithms

Figure 6-7 shows that the MedSOM algorithm maintained a processing speed advantage across all SOM dimensions, MeSH numbers, and article numbers. MedSOM was 5% faster (81 versus 85 minutes) than LibSVM at 350x350, with 32,000 MeSH and 32 million articles, and 5% faster at 350x350 with 32,000 MeSH and 8 million articles. Interestingly, MedSOM's advantage was larger with smaller datasets, taking 1927 milliseconds compared to 2,567 milliseconds at 50x50 with 1,000 MeSH and 8 million articles, a 25% reduction.

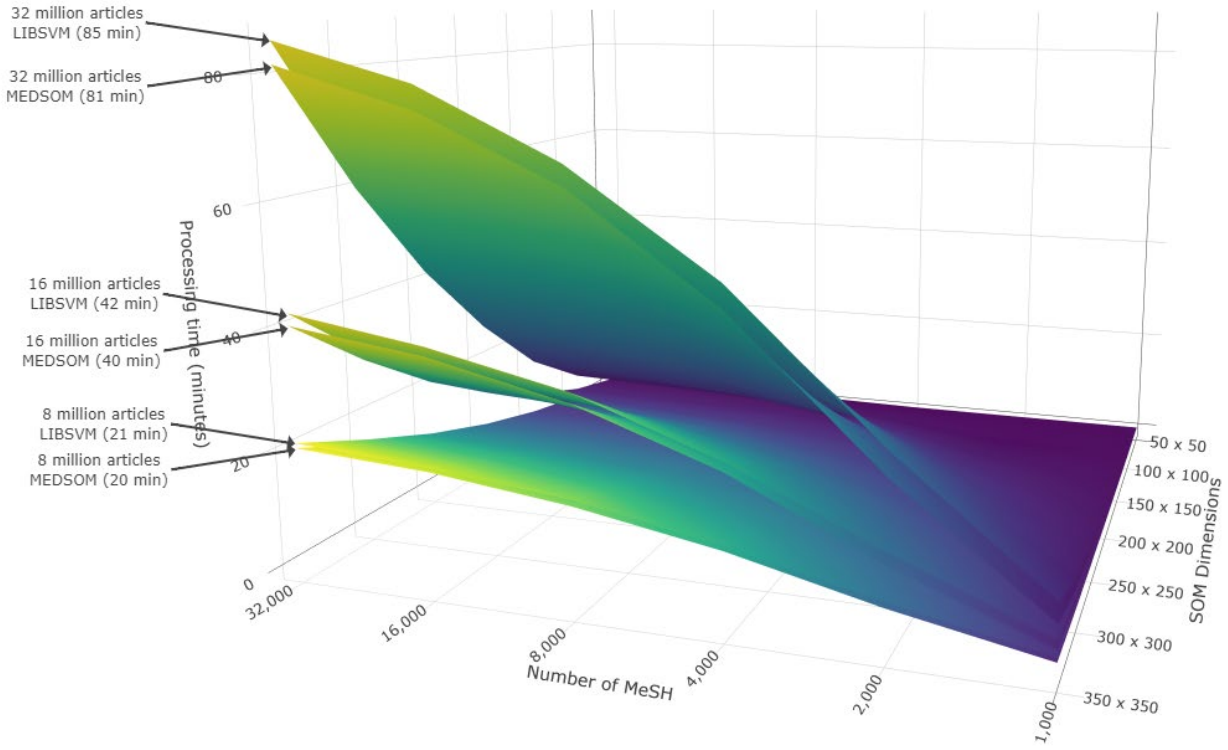


Figure 6-7: Processing time in minutes per cycle for the LibSVM and MedSOM algorithms

6.6 - Discussion

The power of the SOM is to condense and present in human-readable form representations of the large sets of information indexed by the Medline database. Not only does the technical innovation of the MedSOM make it possible to represent the entire database of articles rather than a subset, it makes it possible to represent all the different MeSH that describe different articles. As it is not possible to predict beforehand which patterns may emerge from what subsets of articles and MeSH, a SOM that represents the entire set of knowledge can be assumed to have the best chance of identifying the full set of meaningful patterns contained in the high-dimensional space of the Medline database.

Over and above the order-of-magnitude improvements in memory use and processing time moving from the dense algorithm to the sparse algorithms, MedSOM reduced the memory required to store the training articles by ~50%, and reduced processing time by ~5%. These improvements are important, as

they increase the capacity of the SOM to accommodate the exponentially increasing number of articles indexed by the Medline database. In addition, the 50% reduction in size of the memory used to store articles makes possible further optimizations at the level of the kernels used to implement the SOM algorithms on GPUs.

As discussed in section 3.2, in their current implementation both sparse algorithms are primarily memory bound, particularly because of the use of atomic operations to retrieve article information from general memory. By halving the memory required to represent each article, MedSOM makes it possible to copy twice as many articles to each thread block, eliminating the use of atomic operations, which has the potential to substantially improve processing speed.

In addition, by increasing the capacity of the SOM to incorporate the entire set of information including articles from across the 150-year history of Medline, it becomes possible to do more sophisticated analyses. Most importantly, it becomes feasible to examine changes in the organisation of published medical knowledge over time using techniques such as Denny et al.'s RedSOM.[8]

6.7 - Conclusions

Self-organizing maps are a valuable tool for identifying and visualizing previously unknown but extremely complex relationships in large databases of highly dimensional data. The current research reports a sparse algorithm that allowed for the training of a 350x350 node SOM on the entire dataset of more than 30 million peer reviewed medical articles indexed by the Medline database with the complete set of 29,917 MeSH, compared with a dense algorithm that was limited to a 275x275 node SOM and trained on 2.1 million articles with 2,300 MeSH.

This incremental step in the capacity of the SOM technique to provide an intelligible map of the entire set of medical knowledge contained in the published literature also points towards the next step, which will build upon the MedSOM in order to provide an account of the changes in information over time.

6.8 - References

1. Bornmann L, Haunschild R, Mutz R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanit Soc Sci Commun. 2021 Dec 1;8(1).

2. Kohonen T. Self-organizing maps. 3rd ed. Berlin: Springer; 2001.

3. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, et al. Science of science. Science (1979). 2018;359(6379).

4. Boyack KW, Klavans R. Creation and Analysis of Large-Scale Bibliometric Networks. In: Glänzel W, Moed HF, Schmoch U, Thelwall M, editors. Springer Handbook of Science and Technology Indicators. Cham, Switzerland: Springer Nature; 2019. p. 187–212.

5. Glänzel W, Debackere K. Various aspects of interdisciplinarity in research and how to quantify and measure those. Scientometrics. 2022 Sep 1;127(9):5551–69.

6. Skupin A, Biberstine JR, Börner K. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS One. 2013;8(3).

7. Wijayasekara D, Linda O, Manic M. CAVE-SOM: Immersive Visual Data Mining Using 3D Self-Organizing Maps. In: International Joint Conference on Neural Networks. San Jose, California: IEEE; 2011. p. 2471–8.

8. Denny, Williams GJ, Christen P. Visualizing temporal cluster changes using Relative Density Self-Organizing Maps. Knowl Inf Syst. 2010;25(2):281–302.

9. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Validating the knowledge represented by a self-organizing map with an expert-derived knowledge structure. BMC Med Educ. 2024;24(416).

10. Melka J, Mariage J. Efficient Implementation of Self-Organizing Map for Sparse Input Data. In: Proceedings of the 9th International Joint Conference on Computational Intelligence. Funcha, Madeira, Portugal; 2017. p. 54–63.

11. Melka J, Mariage JJ. Adapting Self-Organizing Map Algorithm to Sparse Data. In: 9th IJCCI 2017: Funchal, Madeira, Portugal (Selected papers). 2017. p. 139–61.

12. National Library of Medicine. MEDLINE Database Home [Internet]. MEDLINE Home. 2021 [cited 2023 May 2]. Available from: https://www.nlm.nih.gov/medline/index.html

13. National Library of Medicine. Medical Subject Headings [Internet]. National Library of Medicine Website. 2024 [cited 2024 Jul 16]. Available from: https://www.nlm.nih.gov/mesh/meshhome.html

14. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli B. Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps. In: 19th World Congress on Medical and Health Informatics [Internet]. Sydney: Australasian Institute of Digital Health; 2023. Available from: https://raw.githubusercontent.com/AndrewAmosJCU/PsychSOM/main/ColorCoded.png

15. Bauer HU, Pawelzik K, Geisel T. A Topographic Product for the Optimization of Self-Organizing Feature Maps. In: Moody JE, Hanson SJ, Lippmann R, editors. Advances in Neural Information Processing Systems 4, [NIPS Conference, Denver, Colorado, USA, December 2-5, 1991] Morgan Kaufmann 1992, ISBN 1-55860-222-4. Denver, Colorado: Morgan Kaufmann; 1991. p. 1141–7.

16. NVIDIA. Nsight Compute - Kernel Profiling Guide [Internet]. Nsight Compute Documentation. 2024 [cited 2024 Jul 31]. Available from: https://docs.nvidia.com/nsight-compute/ProfilingGuide/index.html

Chapter 7: Validating the knowledge represented by a self-organizing map with an expert-derived knowledge structure

Authors:

Andrew James Amos[*1], MB.BS, Kyungmi Lee[2], PhD, Tarun Sen Gupta[1], PhD, Bunmi S. Malau-Aduli[1,3], PhD

* Corresponding Author

1 College of Medicine and Dentistry, James Cook University, Townsville, Australia

2 College of Science and Engineering, James Cook University, Cairns, Australia

3 School of Medicine and Public Health, University of Newcastle, Newcastle, Australia

This chapter describes a novel method of validating the knowledge structures extracted from the peer-reviewed medical literature by machine learning techniques by demonstrating that they provide a coherent and consistent interpretation of the knowledge structures inherent in the expert-selected reference lists of a psychiatric textbook.

7.1 - Abstract

Background

Professionals are reluctant to make use of machine learning results for tasks like curriculum development if they do not understand how the results were generated and what they mean. Visualizations of the peer reviewed medical literature can summarize enormous amounts of information but are difficult to interpret. This article reports the validation of the meaning of a self-organizing map derived from the PubMed index of peer reviewed medical literature by its capacity to coherently summarize the references of a core psychiatric textbook.

Methods

Reference lists from ten editions of *Kaplan and Sadock's Comprehensive Textbook of Psychiatry* were projected onto a self-organizing map trained on Medical Subject Headings annotating the complete set of peer reviewed medical research articles indexed in the PubMed database (MedSOM). K-means clustering was applied to references from every edition to examine the ability of the self-organizing map to coherently summarize the knowledge contained within the textbook.

Results

MedSOM coherently clustered references into six psychiatric knowledge domains across editions (1967-2017). Clustering occurred at the abstract level of broad psychiatric practice including General/adult psychiatry, Child psychiatry, and Administrative psychiatry.

Conclusions

The uptake of visualizations of published medical literature by medical experts for purposes like curriculum development depends upon validation of the meaning of the visualizations. The current research demonstrates that a self-organizing map (MedSOM) can validate the stability and coherence of

the references used to support the knowledge claims of a standard psychiatric textbook, linking the products of machine learning to a widely accepted standard of knowledge.

Keywords

Artificial intelligence; Machine learning; Curriculum development; Scientometrics; Medical education; Explainable AI

7.2 - Background

The selection of content for inclusion in the medical curricula of undergraduate university degrees, as well as the curricula of postgraduate generalist and specialist programs for training doctors, is almost entirely based on expert judgement rather than empirical evidence about the relative importance of different medical skills and knowledge for competence in those areas of practice.[1,2] This is largely due to the enormous set of potentially relevant information from which the topics covered by medical curricula must be selected, the rapid accumulation of new information across diverse topics, and the partially mutual, partially conflicting interests of stakeholder groups including patients, clinicians, and decision-makers.

An example of the reliance on expert judgement for the selection of content is curriculum mapping, which describes a type of structured brainstorming intended to reduce the chance that important topics will be left out of a curriculum. It advises developers to consider the needs of different stakeholders such as students and teachers, and different curriculum purposes such as learning and assessment. By contrast, while we have not found any research which attempts to do this, it is technically possible to select curriculum content on the basis of empirical evidence, defined as systematically gathered and evaluated evidence about quantitatively or qualitatively observable phenomena. An example for the purposes of illustration is a medical curriculum developed to address only the 100 diagnoses associated

with the largest set of costs in a health care system. While it is unlikely that such a curriculum would be acceptable to patients, physicians, or administrators, there may be advantages to a curriculum designed by experts who systematically integrated empirical information such as resource-intensive diagnoses into content selection decisions.

In the absence of the widespread use of empirical information to guide content selection, medical curriculum development largely relies upon expert judgement. There is face validity to having domain experts such as physicians, surgeons, or psychiatrists decide what knowledge and skills are core to their practice. However, basing medical practice on expert judgement alone can have negative consequences, particularly where biases are widely shared. For example, it is now known that the previously common practice of excluding women from patient samples in medical research was based on the erroneous assumption that patterns of health, illness, and response to treatment were common to men and women, leading to widespread sub-optimal treatment of cardiac health and illness in women spanning decades.[3]

While there is no doubt that the peer-reviewed literature itself is subject to biases, the provision of objective sources of evidence summarizing features of the research in each field of medicine and their place within medicine more generally could reduce bias in two ways. First, the least biased expert would surely be the one with the best knowledge of an entire field of research. An objective summary of the available research could highlight for the expert areas they have overlooked, including topics of emerging importance, or areas they have overvalued, such as treatments associated with declining research, and help organize their assessment of the relative importance of different areas. At the worst, it could reassure the expert that they currently have an accurate understanding of the totality of the relevant literature.

Second, empirical models of the research literature can be interrogated for bias in ways that expert judgement cannot. For example, now that it is understood that knowledge of cardiovascular health and disease was biased by the exclusion of women, it may be possible to detect similar biases by analyzing the proportion of patients from particular demographic or clinical groups across fields of research. In some cases, the under-representation of women in clinical samples would be expected (research into the detection of prostate cancer, for example) while in other cases it would indicate the possibility of bias indicating follow-up (research into the thresholds for follow-up on screening for heart disease, for example).

The enormous and growing volume of information about medicine means that empirical models will often need to be presented in visual form to be understood. The field of Machine learning (ML) provides techniques capable of producing empirically derived visualizations that meaningfully summarize the large volume of information and level of complexity of databases of medical knowledge. Medline is a freely available database maintained by the National Library of Medicine (NLM) which lists and describes almost all articles published in reliable peer-reviewed medical journals, including articles as far back as the nineteenth century.[4] Skupin has been particularly creative in applying map-making techniques to visualizations of the Medline database to highlight structural features such as the relative frequency with which medical concepts co-occur in research papers.[5]

Despite their promise, there are significant barriers to the adoption of ML models in medical practice, including for curriculum development. Most importantly, clinicians, educators, and other health professionals resist using ML models if they do not understand how those models work, or what the models mean.[6,7]

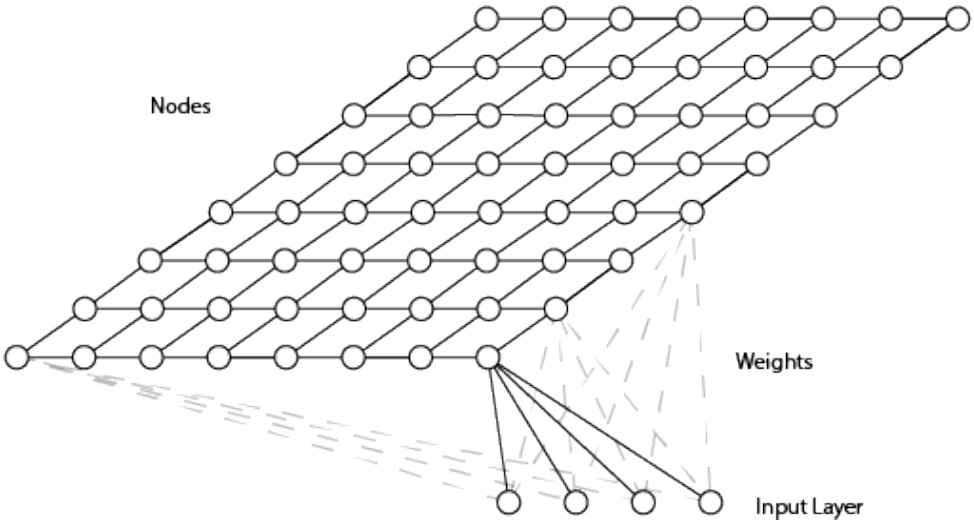
7.2.1 - Visualizing medical knowledge with self-organizing maps

Many ML techniques are used to visualize the knowledge contained within published peer-reviewed scientific literature.[8] Self-organizing maps (SOMs), pioneered by Kohonen, are an attractive method that can be applied to structured or unstructured data, can identify novel patterns not anticipated by researchers, and have successfully projected very high dimensional datasets onto 2D maps while retaining many of the topographical features of the original high dimensional space.[5,9] These features are suited to eliciting, summarizing, and compactly representing the properties of medical knowledge latent in the published literature as an objective source of evidence to guide the selection of content for inclusion in medical curricula. Box 1 presents a summary of the technical features of SOMs.

Box 1: Technical features of Self-organizing maps

Self-organizing maps (SOMs) are a type of neural network composed of a grid of nodes each of which is an independent unit that processes inputs from the environment and which learn over time to recognize specific features of those inputs. To understand SOMs it is necessary to understand: 1) what a SOM does; 2) the representation of information in the SOM; 3) how individual nodes process inputs from the environment; 4) how the whole SOM learns from inputs; and 5) how the SOM represents what it has learned.

Box Figure 6-1: Self-organizing map (8x8 node grid; 4 dimensional input layer; 4 weights per node):



What a SOM does and why this is useful

- A SOM transforms high-dimensional information into a 2-dimensional map which retains many of the properties of the high-dimensional form; In Box Figure 6-1 the four dimensions of the input layer are transformed into the 2 dimensions of the grid of nodes
- In particular, a SOM will retain the property that inputs that were close together in the high-dimensional space will be close together in the 2-dimensional space (topological ordering)
- This is useful because human vision is highly effective and efficient at understanding spatial relationships in 2-dimensions and 3-dimensions, but completely unable to understand higher-dimensional relationships

How information is represented in self-organizing maps

- SOMs represent information as vectors, which are ordered sets of numbers of particular size where the position has meaning; in Box Figure 6-1 the input layer is depicted as a vector of four numbers, and each node receives four weighted inputs from the input layer

- The vectors in this research project are sets of 29,917 numbers where the position represents a specific Medical Subject Heading (MeSH)
- In computer memory, this looks like a set of 1s and 0s: [0, 1, 0, 0…..(x29,917)]
- A vector with a 1/0 in a particular position means the presence/absence of a particular MeSH, for example the vector in the above line might mean: ["Schizophrenia" absent/0, "Bipolar disorder" present/1, "ADHD" absent/0, "Anxiety" absent/0, ….etc]

Representation of articles: The representation of an article is the pattern of 1s and 0s in a vector representing all the MeSH annotating that article (as each article is annotated by an average of 9 MeSH, most of the positions will be 0s)

Vector representing an article (7 numbers shown, 29,910 numbers represented by …):

0	1	0	0	0	…	1	0

Representation of nodes: Each node represents a neuron in the brain, which receives information, processes the information, and produces a level of activation. Each node is specified as a vector of the same length as the inputs (29,917 in this research) where each position in the vector represents a weight which is used to process inputs. Weights are not 0s and 1s, but floating point numbers which vary between certain limits (e.g. between 0.0 and 1.0)

Vector representing a node (7 numbers shown, 29,910 numbers represented by …)::

0.11	0.76	0.01	0.003	0.01	…	0.98	0.001

How individual nodes process inputs from the environment

- In this research, the inputs are articles represented by vectors which specify the MeSH which annotate those articles (as above)

- When an article is presented to the SOM, every node independently processes the information from the vector representing that article
- Each node calculates the distance squared between the weight of the node at a particular position, and the article value at that position

Article

1	0	0	1	0	...	1	0

Vector -

0.11	0.76	0.01	0.003	0.01	...	0.98	0.001

Distance =

0.89 x 0.89	(-0.24) x (-0.24)	(-0.01) x (-0.01)	0.997 x 0.997	(-0.01) x (-0.01)	...	0.02 x 0.02	(-0.001) x (-0.001)

SUM SQUARES = 1.84 [ignoring the other 29,910 MeSH positions]

- The node with the lowest sum of squares after input from a particular article is described as the best-matching unit or BMU; this node is the one where the weight vector is most similar to the article vector

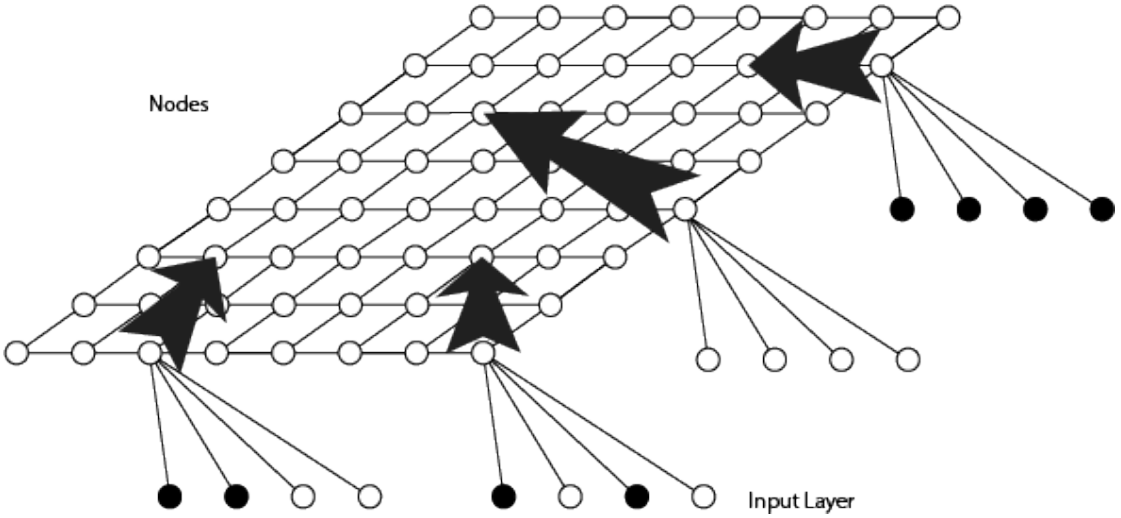
How the whole SOM learns from inputs

- In the most simple case, before training starts the weights of every node in a SOM will be set to random numbers
- As a result, the BMU of each article will be more or less randomly distributed across the grid of nodes, because all article vectors will be equally likely to be similar to all node vectors
- The SOM learns by changing the node weights of the BMU after each presentation of an article by moving those weights closer to the article vector; it also moves the weights of the

nodes immediately surrounding the BMU towards the article vector, but moves them less than the weights of the BMU

- Over multiple iterations of presentation of all articles to the SOM, this process leads to topological ordering of the SOM grid such that inputs that were close together in the high-dimensional space will be close together in the 2-dimensional space (topological ordering; Box Figure 6-2)

Box Figure 2: Learning in Self-organizing map:



Box Figure 6-2 shows four inputs with different vectors learning to differentiate themselves over multiple iterations by activating spatially diverse nodes separated on the face of the grid. Black dots represent true/1 and white dots represent false/0 in the inputs. Over learning the BMU shifts from the original random node to a node which reproduces closeness in the high-dimensional space represented by the input (4-dimensions in Box Figure 6-2) by closeness in the 2-dimensional grid of the SOM.

While far from the only source of useful information about the practice of medicine, the corpus comprising the entire set of peer reviewed literature indexed by databases such as Medline,[4] Web of Science,[10] and Scopus,[11] appears likely to be the most comprehensive and authoritative compendium of evidence based medical knowledge independent of the experiences and biases of individual clinicians. If this premise is accepted, then maps derived using ML methods to summarize and organize that evidence have the potential to provide the most reliable empirical basis with which to guide the selection of content by experts engaged in curriculum development. To be useful for curriculum development, maps of medical knowledge must provide experts with access to readily understandable information about the structure and characteristics of the medical literature that are not otherwise available, and that are consistent with experts' current understanding of the literature.

Evaluation of existing efforts to map the medical literature have focused on technical features such as the accuracy of clustering[12,13] or the visual idioms of presentation[5] more than the meanings conveyed by the maps to experts attempting to integrate them into their own practice. As an example of the latter, if a cardiovascular expert involved in curriculum development became aware that women had been excluded from earlier cardiac research, it would be useful for them to be able to consult a map of the peer-reviewed research about the treatment of heart disease which color-coded the representation of other groups previously under-represented in medical research. Such a map would make the distribution of adequate- versus under-representation of specific groups in specific areas of research intelligible to the expert, who could then draw on their expertise to decide what use to make of that information. The current research is a necessary step towards this type of map.

7.2.2 - Integrating self-organizing maps into the curriculum development process

Thus, SOMs are powerful analytic tools with the potential to facilitate curriculum development by condensing enormous sets of information into 2-dimensional maps which highlight relevant information using visual cues like color. Translating this potential into practice requires several steps. The first step is to show it is technically possible to model the entire Medline database in a single SOM. Amos et al. (2023) used SOMs to visualize the entire corpus of peer-reviewed medical literature indexed by the database Medline.[14] Amos et al (in preparation) extended their original SOM using the Relative Density SOMs (ReDSOMs) technique developed by Denny et al. (2010) to show how the knowledge indexed by Medline evolved over time.[15]

Amos et al (2022) further extended the approach by identifying psychiatric topics of emerging importance in the Medline database between 1972 and 2016.[16] Also planned is an experimental test of whether the emerging topics identified by Amos et al. (2022) are perceived to be useful for the purpose of planning professional development sessions by a group of psychiatrists.

The current research describes a novel method of validating the SOM developed by Amos et al. (2023) by examining its ability to meaningfully interpret the knowledge represented by the expert-selected references across the editions of a psychiatric textbook. The ultimate goal is to develop SOMs in forms useful for integration into curriculum development or curriculum maintenance activity, for example by highlighting emerging topics for consideration for inclusion in a curriculum; or by highlighting topics of declining importance for consideration for removal.

7.2.3 - Study Context

The ML field is starting to realise that the understandability and acceptability of its models is the main factor in determining whether they will be used by experts.[17,18,19] Evaluations of current ML models of the medical literature generally focus on technical features such as accuracy rather than indicators of

understandability and acceptability such as the meaning of the models to end-users. We developed a novel approach to the validation of a SOM of the medical literature by examining whether it meaningfully organized the implicit knowledge structures represented by reference lists of different editions of a psychiatric textbook.

For the purposes of this research an explicit knowledge structure in a set of articles is any relationship between two or more articles likely to be immediately obvious to a human reader. For example, a human reader will immediately recognize the grouping of articles by the journal in which they are published. An implicit knowledge structure is any pattern of relationships between articles that is unlikely to be readily apparent to a human reader, but is detectable by an ML technique like a SOM. An example of an implicit knowledge structure is a bias against the inclusion of female participants in a particular type of treatment trial that is revealed by the visual properties of a SOM.

7.2.4 - Expert-derived knowledge structure

Kaplan and Sadock's Comprehensive Textbook of Psychiatry is the dominant textbook for the study of psychiatry in the US and elsewhere (afterwards referred to as *KSCTP*).[20] First published in 1967, the tenth edition arrived in 2017,[21] with new editions appearing somewhat irregularly but generally close to 5 years apart. Usefully for the validation of a map of the medical literature, *KSCTP* has included detailed reference lists in support of its knowledge claims starting with the first edition. We hypothesized that a SOM of medical knowledge trained on the MeSH of the complete set of Medline articles would meaningfully interpret the knowledge structures implicit in the references published within each edition of *KSCTP*. For example, we expected that the complete set of references within each edition of *KSCTP* would be represented by specific regions of the SOM which would change in a coherent and understandable way over time.

7.2.5 - Purpose of the study

This paper aims to demonstrate that SOMs can meaningfully visualize features of psychiatric knowledge contained within the peer reviewed literature consistent with the organization implicit in the reference lists of the independently developed expert-derived textbook *KSCTP*. In previous work the same authors applied a SOM of 350 x 350 nodes to the complete set of articles indexed by the Medline database to show the relative importance, relationships between, and evolution of, domains of medical knowledge represented by Medical Subject Headings (MeSH) defined by the NLM;[22] and used a custom machine-learning algorithm developed by Ohniwa et al.[23] (the incremental statistic) to identify topics of emerging importance in the psychiatric subset of the Medline indexed literature.[24] The current paper is designed to investigate the extent to which the relationships extracted from the Medline database by SOMs are consistent with the organization of psychiatric knowledge across ten editions of a core psychiatric textbook, answering the research questions:

1. To what extent is a SOM trained to extract the implicit organizational structures of medical research indexed by the Medline database consistent with the implicit expert-derived organizational structures of a core psychiatric textbook?
2. How does the interpretation of the psychiatric knowledge represented by each edition of the textbook provided by its projection onto the SOM model change across those editions?

7.3 - Methods

7.3.1 - Mapping the knowledge covered by a psychiatric textbook

The changes between editions of the core psychiatric textbook *Kaplan and Sadock's Comprehensive Textbook of Psychiatry* were visualized by projecting the articles contained within the reference lists of the entire textbook:

- A complete set of references from all editions of the textbook was projected onto the static SOM from Amos et al (2023)
- Reference sets from each edition of the textbook were projected onto the SOM

7.3.2 - Training the Self-Organizing Map of the Medical Literature (MedSOM)

As we have described in more detail elsewhere,[14] the SOM for this research comprised a set of nodes arranged in a square 2D matrix (350 x 350 nodes). This shape was chosen during the previous research by experimentation with different sizes, starting with the 275x275 nodes of an earlier model which was trained on a smaller subset of the Medline database than ours, and increasing up to 400x400 nodes by adding 25 nodes in each dimension. With increasing size the topographic error of the model first declined and then reached a plateau at 350x350 nodes, leading us to select this size of map for the current research.

The training set included the entire set of 33,375,863 peer-reviewed articles published by the NLM in its Medline database as of 1.1.2022. Each article was represented as a binary vector with 29,917 elements encoding the presence or absence of each of the 29,917 Medical Subject Headings (MeSH) in the set published by the NLM as of 1.1.2022, after excluding administrative codes and MeSH annotating less than 100 articles.

The MeSH are a set of phrases with defined meanings maintained as a controlled vocabulary by the NLM. They are organized in a tree-like hierarchy with categories such as "Anatomy", "Diseases", and "Disciplines and Occupations" at the top, with more and more specific categories at lower levels. For example, one of the pathways under "Diseases" narrows its meaning through the phrases "Infections",

"Bacterial infections", and "Bacterial Zoonoses" at the lowest level. The NLM controls the meaning of each MeSH and their organization into a hierarchy. They continuously review the peer-reviewed medical literature for new phrases to be defined and added to their controlled vocabulary and its hierarchy. One of the top-level categories is "Psychiatry and Psychology".[24]

On average, each of the 33 million articles had been annotated with 9 MeSH describing its main features. For example, an article describing a clinical trial of high-dosage haloperidol for patients with chronic schizophrenia would be annotated with MeSH representing: "Adult population, Human", "Schizophrenia", "Haloperidol", among others. For SOM training purposes, this article would be represented by a vector of 29,917 binary elements in which ~29908 would be false/0 (indicating that the MeSH did not describe the article) and ~9 would be true/1 (indicating that the MeSH did describe the article – see Box 1).

Kohonen's batch training algorithm was implemented using sparse matrices.[9,14,25,26] In each epoch, every article was presented to the SOM, which calculated the best matching unit and the second best matching unit, with error accumulated over all articles and weight changes applied at the end of the epoch, using the algorithm described in Table 7-1.

Table 7-1: Training algorithm – Parallel Batch SOM with sparse binary matrices

Input: *x*: *N* sparse vectors of *D* elements representing Medline articles
Data: *w*: initialized codebook of *M* dense vectors of length D; represents weights between inputs and each SOM node
Data: χ: array of *N* reals; as *x* is sparse this matrix is the number of non-zero elements per row (i.e. the number of MeSH annotating each article)
Data: *dst* array of *N* reals storing distance to best matching unit for each article

Data: *bmu1, bmu2* – 2 vectors of *N* integers to store best matching and second best matching unit/node for each article

Data: *num* – array of *D* reals to accumulate numerator values

1 **Randomise** codebook weights *w* between 0.0f and 1.0f

2 **for** each epoch $e \leftarrow 1$ to K **do**

3 **compute** **σ**€; // radius for current epoch

4 **standardise** codebook weights; // sqrt of sum of weights per node

5 **calculate** bmu1, bmu2 for each article $i \leftarrow 1$ to N

6 **for** each node $m \leftarrow 1$ to M

7 **sum** $m \cdot i$ // dot-product node weights · article MeSH

8 **calculate** $dst = -_a - 2 \cdot (m \cdot i)$

9 **calculate** new weights for each node ← 1 to M **do**

10 **for** each article $i \leftarrow 1$ to N

11 **calculate** $I(e) = \exp(-||r_k - r_c||^2/(2\ \boldsymbol{\sigma}(e)^2)$ // Neighborhood fn

12 **reduce** denominator = accumulate h_{ck} per node

13 **calculate** numerators = accumulate h_{ck} per node/article

14 **calculate** new weights = numerators / denominator

15 for $i \leftarrow 1$ to N

16 **calculate** adjacency = (bmu1-bmu2 =< 1) // Manhattan distance

17 **calculate** topographic error = % adjacent

Note: D - total number of MeSH categories (=29,917–; D_a - number of MeSH categories annotating an individual article; N - number of articles (=33,375,866); M - number of nodes (350x350=122,500); –(e) - width of the neighborhood (which changes over training epochs according to the formula $175/(1.7)^{epoch}$)

Twenty training epochs were completed, with calculation of the topographic error after each epoch. The codebook with the lowest topographic error was selected for further analysis. The nodes comprising the MedSOM represented by this codebook were divided into those representing relatively psychiatric knowledge and relatively non-psychiatric knowledge. Nodes where the input weights gave a higher priority to psychiatric than non-psychiatric MeSH were categorized as psychiatric, with all others categorized as non-psychiatric. Psychiatric MeSH are those organized as sub-categories under the top-level "Psychiatry and Psychology" category in the NLM's controlled vocabulary hierarchy.

7.3.3 - Projecting textbook editions onto a published medical literature map

Each edition of the core psychiatric textbook *KSCTP* reports detailed reference lists comprising the scientific evidence base for the asserted knowledge claims. The citations include books, peer reviewed articles, and grey literature including government reports, web pages, and other sources. Complete reference lists were obtained from each of the 10 editions published before 2023, and the R programming package easyPubMed was used to retrieve the unique PubMed ID (PMID) and list of Medical Subject Headings (MeSH) describing each referenced article.

Table 7-2 – Organizational features of *Kaplan & Sadock*[21] textbook across editions

Edition	Year	Pages	Chapters	New Chapters	Removed chapters	Citations	Citations per chapter	Pubmed citations	% Persisting citations
1st	1967	1629	53	53	N/A	2927	55.2	585	N/A
2nd	1975	2572	52	18	19	9126	175.5	2280	44%
3rd	1980	3306	57	9	4	17030	298.8	4225	56%

4th	1985	2055	54	3	6	2091	38.7	407	5%
5th	1989	2158	50	10	14	3999	80.0	780	21%
6th	1995	2805	53	8	5	8625	162.7	2393	16%
7th	2000	3345	55	7	5	10180	185.1	4257	23%
8th	2005	4064	55	2	2	11795	214.5	3780	20%
9th	2009	4521	59	4	0	10472	177.5	4815	25%
10th	2017	4533	62	5	2	9235	149.0	4157	26%

Table 7-2 describes the structural features of the *KSCTP* textbook.[21] It maintained a standard edited format across all editions, with a low of 50 chapters in 1989 (5th edition) and a high of 62 chapters in 2017 (10th edition). Single or multiple invited authors were responsible for subsections of each chapter, and the number of subsections per chapter varied from a low of 4.39 in 1967 (1st edition) to a high of 6.85 in 2005 (8th edition). There was an average of 7.3 new chapters added to each edition, with an average of 6.3 chapters either removed or merged, with more changes in earlier editions. On average, 26% of PubMed citations persisted from one edition to the next, with a low of 5% in 1985 (4th edition) and a high of 55% in 1980 (3rd edition).

The complete text of each reference was submitted to the Medline search function using the get_pubmed_ids function within the r package *easyPubMed*, and the complete xml of each match was retrieved using the easyPubMed::fetch_pubmed_data function. The accuracy of retrieval was confirmed by manually checking that the first author, title, and journal specified in the reference text matched those fields in the retrieved information. Non-matching references were discarded, and the number of references in each edition confirmed/not confirmed to be indexed by the Medline database were recorded. As only references which were matched in the Medline database were retained, only Medline-

indexed articles were included in the analysis. Records were kept of both the number of Medline-indexed references and the total number of references for each edition.

7.3.4 - Interpreting the meaning of the projection of textbook knowledge onto MedSOM

MedSOM is a 2-dimensional representation of the organization of all the information contained within the Medline database, comprising 350x350 = 122,500 nodes laid out in a square grid (see Box 1 for a description of the technical features of SOMs). Each of the individual nodes is characterized by a set of 29,917 weights which each represent one of the 29,917 MeSH used by the NLM to annotate articles. Each article was represented by a vector of 29,917 boolean variables corresponding to the node weights and MeSH. For each boolean a true value indicated that an article was annotated with a specific MeSH, while a false value indicated the article was not annotated with that MeSH. Just as each article was annotated with on average 9 MeSH, so the representation of the article in the SOM was a boolean vector with on average 9 true and 29,908 false, boolean variables. After training, nodes close to each other in the SOM had similar MeSH weights, and therefore were more likely to be the best matching unit for articles close together in the high-dimensional space represented by the annotating MeSH.[9]

For the purposes of providing an external validation of the organizational structure inferred by the SOM it is necessary to show that it can coherently organize the knowledge structures implicit in some other set of information. In the current study, that set of information is the complete set of articles referenced across all editions of the *KSCTP* textbook, and within each edition.

The list of MeSH describing each retrieved article was input to the MedSOM to identify the node best representing that article. The structure of knowledge represented by these articles was then projected onto MedSOM as a single set containing all editions, and individually for each edition. K-means clustering was applied to the position of each article on the MedSOM grid for the complete set and

individual edition-sets of articles to identify groups of related articles, using the kmeans function in the r *stats* package.

For each of the complete set and individual edition-sets of articles the optimal number of clusters was determined using the r *stats* package's *fviz_nbclust* function using the "elbow" method. The elbow method seeks to minimize both the number of clusters and the sum of squares of the distance between articles within each cluster by finding the cluster number beyond which there is a relative plateau in further reductions of sums of squares.

The success or failure of the SOM to provide a meaningful interpretation of the expert-derived organizational structures of the *KSCTP* textbook can be judged by the coherence of the interpretation provided for the entire set of references across all editions, and its persistence between editions. Before training, the SOM would be expected to distribute articles at random across its nodes. After successful training, it would cluster articles covering similar topics close together, both within each edition, and across all editions.

The meaning of the knowledge clusters was explored by extracting the most common MeSH annotating articles in each cluster and identifying the 10 articles in the cluster with the largest number of those common MeSH.

7.4 - Results

Figure 7-1 shows the total number of references in the reference lists of each edition of *KSCTP*, alongside the number of those references indexed by the Medline database. As the Medline database only indexes scientific articles published in peer-reviewed medical journals, the difference between total and Medline indexed references in each edition is made up of non-indexed citations such as books, newspaper articles, and grey literature.

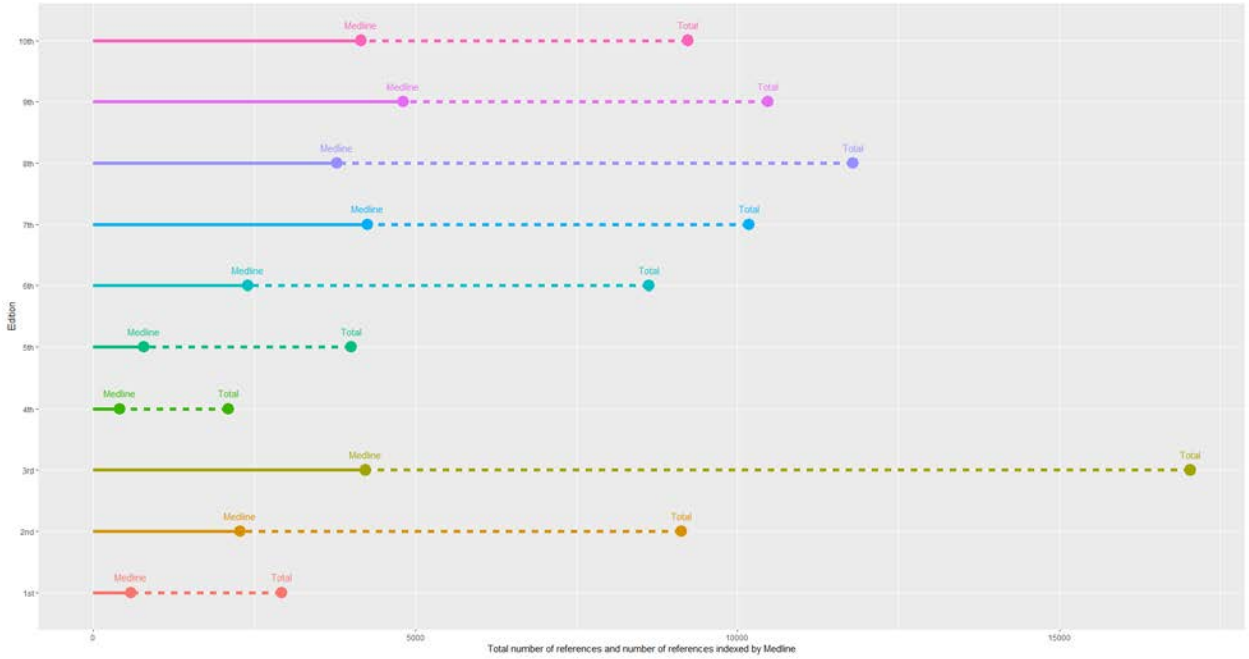


Figure 7-1. Number of references by edition (Medline indexed references and Total references indexed by edition)

The discontinuity after the third edition with a large reduction in the total and indexed references was due to a conscious decision by the editors of the *KSCTP* textbook. The rationale provided in the fourth edition foreword was that the rapidly escalating number of references over the first three editions had achieved the desired result of establishing the authoritative nature of the textbook grounded in the published literature, allowing for more selective citations in future editions.[27] From the very low baseline of the fourth edition there was a steady increase in the number of total references and references indexed in Medline until the 8th edition, after which numbers stabilized at close to 10,000 total references and close to 5,000 Medline indexed references. Thus, in later editions of *KSCTP*, about half of the references are indexed, peer-reviewed articles, a significant change from the 3rd edition (published

1980) where less than one quarter of the references were indexed, peer-reviewed articles, with a much higher reliance upon books.

7.5 - Consistency, coherence, and meaning of SOM projections

The purpose of this research is to show that the SOM trained on all Medline articles and MeSH provides a meaningful, coherent, and consistent interpretation of the knowledge contained within the *KSCTP*. In this exploratory phase, coherence is demonstrated by an understandable organization of the knowledge into identifiable categories; consistency is demonstrated by the persistence of the same identifiable categories over different editions of the *KSCTP*; and meaning is inferred by examination of the qualities of the articles clustered together on MedSOM. Coherence and consistency are demonstrated in Figures 7-2 and 7-3, while meaning is inferred from Tables 8-3 and 8-4. While future phases of research will focus on optimizing formal properties of the SOM, that is not the goal of this phase of research.

Figure 7-2 shows the results of the projection of all articles referenced by *KSCTP* onto the MedSOM, then subject to k-means cluster analyses, where each separate cluster is color-coded. The first facet shows the projection and clustering of all references from all editions, and the other facets each show the results using references from a single edition. The six clusters describe coherent and separate domains of psychiatric knowledge which are stable across the editions. The MedSOM does show variations consistent with the underlying data set, for example by a rapid increase in the density of articles across the first three editions, and then a steady increase from the fourth to eighth editions followed by a plateau.

Figure 7-3 reproduces the facets of Figure 7-2 along the diagonal cells and shows the superimposition of all the articles in each edition on all the articles of each of the other editions. The smaller size of each cell of the matrix in Figure 7-3 made the points in the scatter plots of Figure 7-2 difficult to compare

across editions. The *stat_ellipse* function of the R *ggplot2* package was used to superimpose a shaded normal ellipse containing 95% of the articles in each domain using the same color coding. There is a high degree of overlap between the position of articles across editions. In order to show where there is overlap and where there are differences between editions, the cells beneath the diagonal show the editions labelled on the y-axis in yellow above the editions labelled in brown on the x-axis. The cells above the diagonal reverse this pattern.

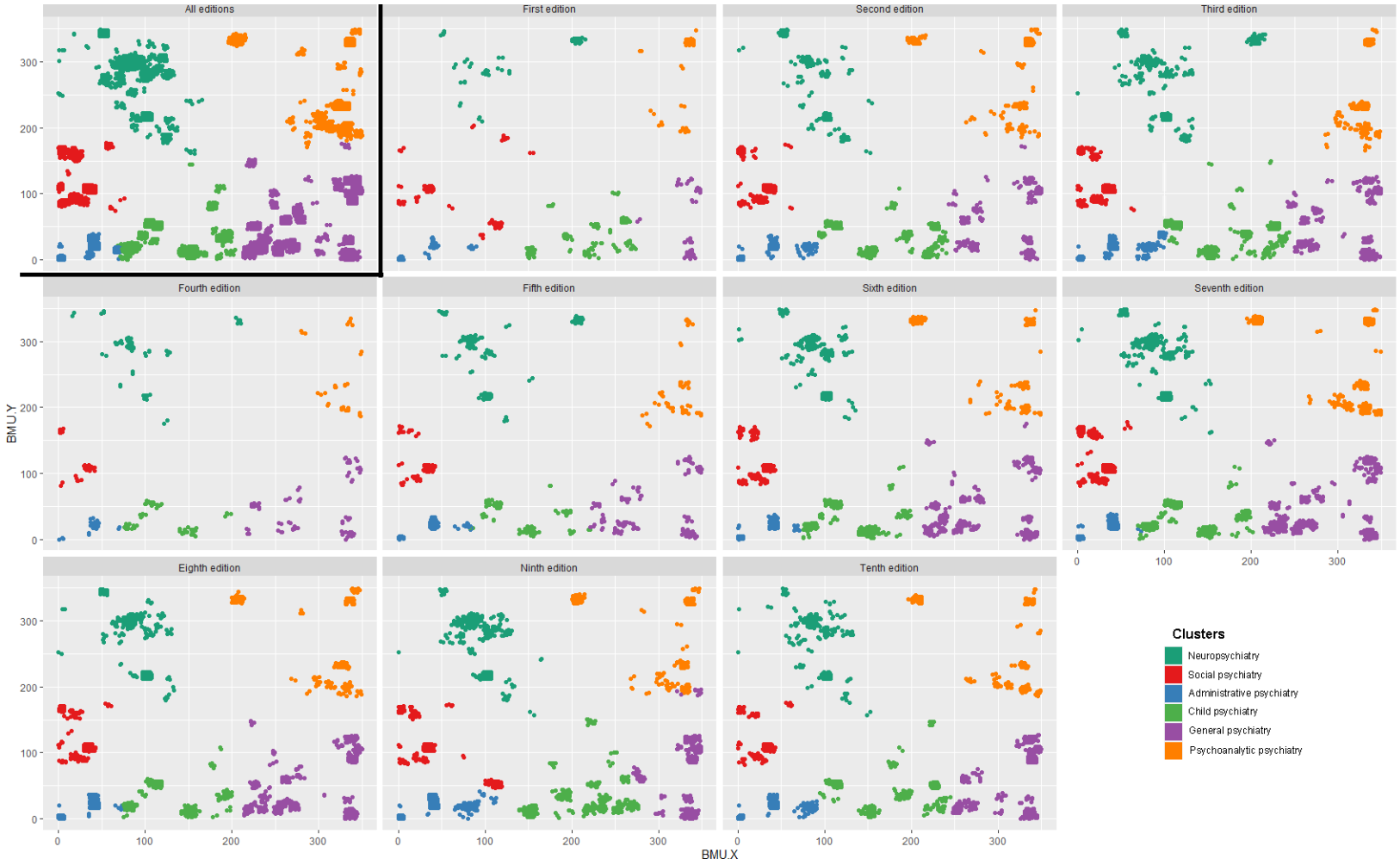


Figure 7-2. Knowledge clusters by edition

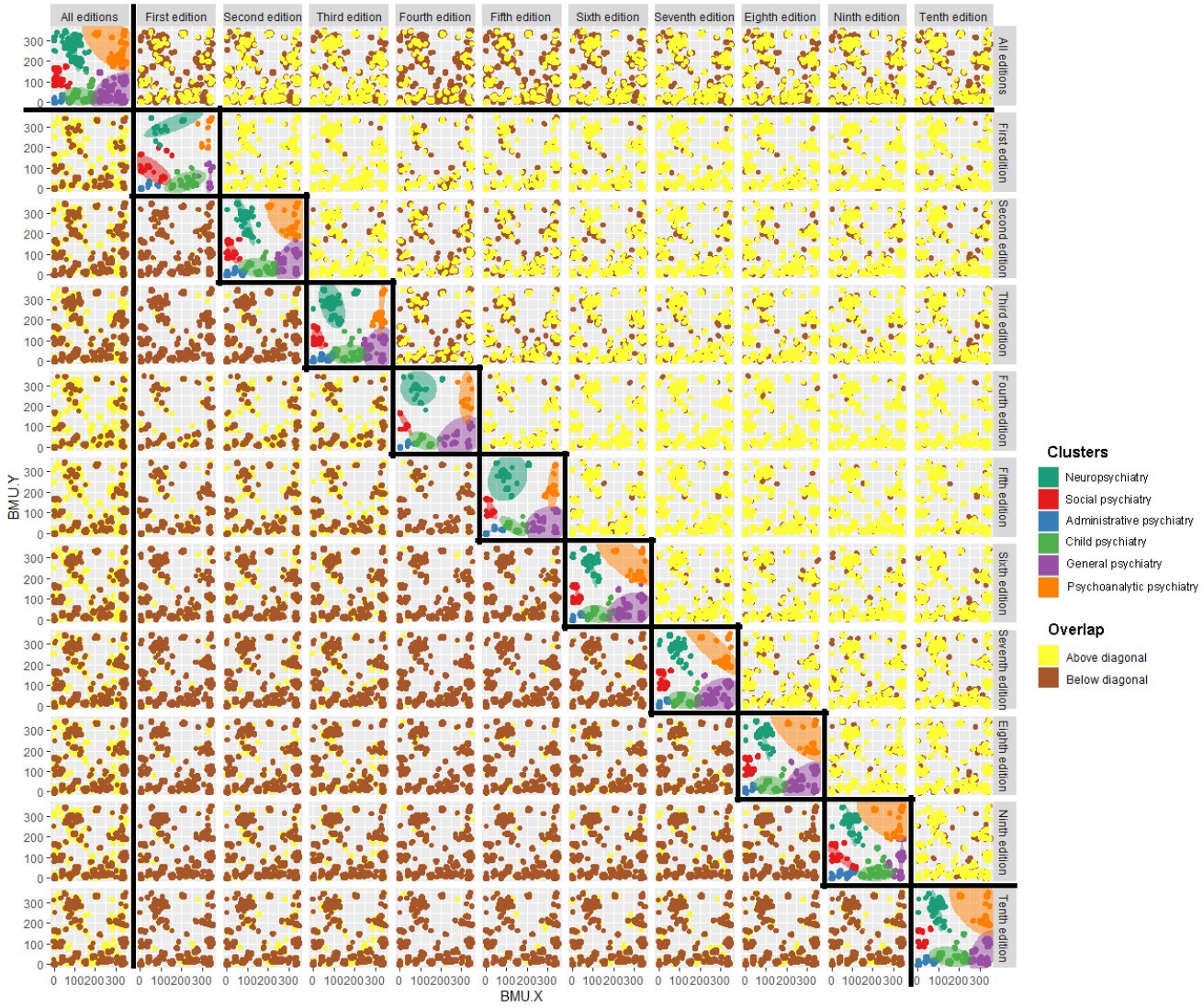


Figure 7-3. Knowledge cluster overlap - overall and by edition

Beneath the diagonal the first column shows where each individual edition overlaps the complete set. As expected, the first edition does not cover much of the content of the complete set of information, but the second and third editions rapidly increase coverage. While the third edition has the greatest number of references and therefore covers most of the points of the complete set, it least well covers the Neuropsychiatry and Psychoanalytic Psychiatry sections. Conversely, the tenth edition covers all clusters relatively well while also containing areas of non-overlap across all clusters.

The third edition pattern is the result of changes in the knowledge base and practice of psychiatry around 1980, the year of publication of the third edition, as well as the third edition of the Diagnostic and Statistical Manual of Mental Disorders, which radically changed the nosological approach of mainstream psychiatry.[28] The third edition pattern reflects an increasing focus on the basic sciences underlying mental disease and a decreasing focus on psychoanalytic psychotherapy. The tenth edition pattern is consistent with the stabilization of the psychiatric knowledge base with the non-overlapping sections representing all the areas of knowledge and clinical practice which have fallen out of use since 1967.

Another factor likely to have influenced this pattern is the rapid increase in publication rates over the decades since 1967. The NLM reports that Medline registered 186,843 citations in 1967 and 986,012 citations in 2021.[29] As the MedSOM was trained on all articles indexed by Medline, the structures of knowledge present in 1967 (when the 1st edition of *KSCTP* was published) may have less impact on the MedSOM than structures in 2017 (when the 10th edition was published).

Table 7-3. Cluster-defining Medical Subject Headings

Cluster MeSH	Neuropsychiatry	Social psychiatry	Administrative psychiatry	Child psychiatry	General/adult psychiatry	Psychoanalytic psychotherapy
1	Oxygen Radioisotopes	Beneficence	False Positive Reactions	Appetite Regulation	Adult	Depressive Disorder
2	Regional Blood Flow	Ethical Theory	False Negative Reactions	Behavior	Mental Disorders	Dopamine
3	Tomography, Emission-Computed	History, 20th Century	Psychiatric Status Rating Scales	Child	Middle Aged	Latency Period, Psychological
4	Behavior, Addictive	Personal Autonomy	Mental Disorders	Child, Preschool	Shame	Psychosexual Development
5	Circadian Rhythm	Physician-Patient Relations	Diagnosis, Differential	Conditioning, Operant	Hospitalization	Ego
6	Cues	Behavior Therapy	Psychometrics	Enuresis	Psychiatric Status Rating Scales	Electroconvulsive Therapy
7	Memory	Bipolar Disorder	Schizophrenia	Food	Attitude to Health	Norepinephrine
8	Nervous System Physiological Phenomena	Continuity of Patient Care	Sensitivity and Specificity	Infant	Borderline Personality Disorder	Psychotherapy, Group
9	Neuropsychological Tests	Ethical Analysis	Antipsychotic Agents	Migraine Disorders	Diagnosis, Differential	Depression
10	Alzheimer Disease	Euthanasia, Active	Borderline Personality Disorder	Weight Loss	Guilt	Individuation

* Medical Subject Headings (MeSH) associated with a psychiatric diagnosis indicated in **Bold.**

Table 7-3 reports the 10 most common MeSH characterizing each of the 6 clusters found for each individual edition and for the complete set. Table 7-4 collects articles with the largest number of common elements for each cluster across all editions. These tables describe the meaning of the organization of knowledge extracted from the Medline database by the MedSOM. For example, articles in the top left Neuropsychiatry cluster are labelled with MeSH related to neuroscience, particularly neuroimaging terms such as Oxygen Radioisotopes and Regional Blood Flow, while articles in the lower-left Social Psychiatry cluster are labelled with social scientific MeSH such as Beneficence and Ethical Theory.

A notable feature of the pattern of MeSH most characteristic of each cluster is the relative absence of diagnostic labels. While Table 7-3 does include a small number of diagnoses, such as Bipolar Disorder in the Social Psychiatry and Schizophrenia/Borderline Personality Disorder in the Administrative Psychiatry cluster, they appear to be grouped based on features other than their clinical manifestations. For example, the inclusion of Schizophrenia and Borderline Personality Disorder in the Administrative Psychiatry cluster can be explained as the result of their importance in epidemiological studies focused on differential diagnosis rather than symptoms or treatment.

Table 7-4. Best matching articles by Cluster

Cluster	Article titles
Neuropsychiatry	A biologist examines the mind and behavior. **Psychosurgery today: psychiatric aspects.** Serotonin, cerebral blood flow, and cerebral metabolic rate in geriatric major depression and normal aging. **Changes in regional cerebral blood flow elicited by craving memories in abstinent opiate-dependent subjects.** Frontal lobotomy in early schizophrenia. Long follow-up in 415 cases. **The dementia of dementia praecox.** Effect of schizophrenia on frontotemporal activity during word encoding and recognition: a PET cerebral blood flow study. **Sertraline. A review of its pharmacodynamic and pharmacokinetic properties, and therapeutic potential in depression and obsessive-compulsive disorder.** Limbic activation during cue-induced cocaine craving. **Mechanisms of lithium action.** An evaluation of bimedial leucotomy. **Estrogen-serotonin interactions: implications for affective regulation.** Initial masking of organic brain changes by psychic symptoms: clinical and electroencephalographic studies. **A long term follow-up of schizophrenics treated with regressive ECT.** Transduction of psychosocial stress into the neurobiology of recurrent affective disorder **Dopamine receptor binding predicts clinical and pharmacological potencies of antischizophrenic drugs**.
Social psychiatry	The ethics of care and treatment of sex offenders **Three types of peer tutoring: effects on the attitudes of students with learning disabilities and their regular class peers.** Comorbidity of personality disorders and depression: implications for treatment. **Longitudinal patterns of anxiety from childhood to adulthood: the Great Smoky Mountains Study** Empathy: misconceptions and misuses in psychotherapy. **The psychotherapeutic utility of the five-factor model of personality: a clinician's experience.** Escalation of aggression: experimental studies. **Substance abuse and adolescent suicidal behavior.** Demonstrating translational research for mental health services: an example from stigma research. **An improved detoxification technique for heroin addicts.** Dealing with our losses. **Group therapies for nursing home adults: an evaluation of two treatment approaches.** Empirically supported treatments for children with phobic and anxiety disorders: current status.

	Increased depressive ratings in patients with hepatitis C receiving interferon-alpha-based immunotherapy are related to interferon-alpha-induced changes in the serotonergic system. Cognitive-behavioral treatment of school phobia. **Preliminary report on the application of contingent reinforcement procedures (token economy) on a "chronic" psychiatric ward.** AMERICAN PSYCHIATRY AND THE CRIMINAL: A HISTORICAL REVIEW. **The reinforcement of behavior in institutional settings.** The effects of social skills training and peer involvement on the social adjustment of preadolescents. **The consequences of open and closed adoption for older children.** Clinical considerations in group treatment of narcissistic disorders. **Behavior therapy and sex therapy.**
Administrative psychiatry	Citizen participation in the development of acommunity mental health center. **The right to refuse treatment with antipsychotic medications: retrospect and prospect.** Field trial for autistic disorder in DSM-IV. **Gaps in doctor-patient communication. Patients' response to medical advice.** The physician-elderly patient-companion triad in the medical encounter: the development of a conceptual framework and research agenda. **Discussion of medical errors in morbidity and mortality conferences.** Performance of screening and diagnostic tests. Application of receiver operating characteristic analysis. **Treating substance-use disorders among physicians.** Limitations of listing specific medical interventions in advance directives. **On wearing two hats: role conflict in serving as both psychotherapist and expert witness.** The ethics of therapeutic modality choice. **Routine laboratory testing for medical disorders in psychiatric inpatients.** Pharmaceutical care role model in psychiatry--pharmacist prescribing. **Application of the predictive value model in the analysis of test effectiveness.** Death due to treatment. **Predictive validity of certification by the American Board of Internal Medicine.**
Child psychiatry	Ontogenetic development of the human sleep-dream cycle. **Incontinence and enuresis.** Comorbidity of parental anxiety disorders as risk for childhood-onset anxiety in inhibited children. **The effectiveness of group psychotherapy with children.** Short-term group psychotherapy for children: an overview. **Anxiety sensitivity and panic disorder.** Tests of competency to consent to treatment. **Developmental dyscalculia: a brief report on four cases.** The borderline diagnosis in adolescents: symptoms and developmental history. **The adolescent and the "hidden" parent.** A follow-up report on children who had atypical sexual experience.

	A general test of motor impairment for children. Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception.
Adult Psychiatry	Schizophrenia in the National Academy of Sciences-National Research Council Twin Registry: a 16-year update. **Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia.** Patient refusal of hydration and nutrition. An alternative to physician-assisted suicide or voluntary active euthanasia. **Alcohol and temporal lobe dysfunction. Some of its psychomotor equivalents.** Defeminization and adult psychological well-being among male homosexuals. **Commonly prescribed medications and potential false-positive urine drug screens.** Shame and humiliation in the medical encounter. **Detection, prevention and retardation of menopausal osteoporosis.** Unmet service needs in methadone maintenance. **Overview: the "wife-beater's wife" reconsidered.** Morbidity following sudden and unexpected bereavement. **The value of psychiatric treatment: its efficacy in severe mental disorders.** DEPRESSION AMONG MEDICALLY ILL PATIENTS. **The irritable bowel syndrome. A clinical review and ethical considerations.** Predictors of posttraumatic stress disorder and symptoms in adults: a meta-analysis. **THERAPEUTIC EFFICACY OF ANTIDEPRESSANT DRUGS. A REVIEW.** Psychoendocrine research on sexual orientation. Current status and future options. **Efficacy of combinations of intramuscular antipsychotics and sedative-hypnotics for control of psychotic agitation.** Neuroleptic-associated tardive syndromes. **Contemporary conversion reactions : a clinical study.** Psychiatric observations under severe chronic stress.
Psychoanalytic psychiatry	Comorbidity of personality disorders and depression: implications for treatment. **A biologist examines the mind and behavior.** Etiological factors in female transsexualism: a first approximation. **Social influences on "self-stimulatory" behavior: analysis and treatment application.** The psychotherapeutic utility of the five-factor model of personality: a clinician's experience. **How effective are interventions with caregivers? An updated meta-analysis.** Personality disorders and treatment outcome in the NIMH Treatment of Depression Collaborative Research Program. **Thumb and finger sucking.** A self-control behavior therapy program for depression. **A verbal group technique for ego-disturbed children: action to words.** Behavior therapy and sex therapy. **Child care and attachment: a new frontier the second time around.** Activity-interview group psychotherapy: theory, principles, and practice. **Psychotherapy of borderline psychotic children.** A study of Fairbairn's theory of schizoid reactions.

	Contemporary conversion reactions : a clinical study. Group therapy for schizophrenia: a practical approach. **Clinical considerations in group treatment of narcissistic disorders.** A developmental-genetic analysis of common fears from early adolescence to early adulthood.

7.6 - Discussion

The relatively opaque information provided by many ML visualizations of scientific knowledge suggests highlights that the validity of maps of scientific knowledge depends to a large degree on how easily users can perceive the meaningful patterns they summarize.[30] The current research demonstrates that the semantic structures implicit within the Medline database of peer-reviewed medical literature and made explicit by the MedSOM coherently organize the peer reviewed articles cited as references in a core psychiatric textbook. The organizing structure is consistent across all editions and within individual editions, with variations over time that align with actual changes in the clinical and scientific framework underpinning psychiatric practice in the decades between 1967 and 2017.

Comparing the projection of references from different editions of *KSCTP* shows that MedSOM effectively clusters groups of articles with similar content close together, and those with less similar content further apart. Considering the most common MeSH and actual articles clustered together on the MedSOM provides an understanding of the meaning of the underlying 2D distribution of the map. For example, the growth of the Neuropsychiatry domain and the decline of the Psychoanalytic Psychotherapy paradigm is reflected by changes in the extent of overlap of those clusters leading up to the pivotal year 1980.

One of the more useful features of the SOM approach to mapping of the medical literature revealed by our findings is that it identifies previously implicit structures of knowledge linking sets of articles while simultaneously identifying the categories by which they are linked. For example, the articles grouped close together in the Administrative Psychiatry cluster cover a very broad range of clinical and service situations. The associated MeSH are dominated by technical concepts such as False positive/False negative reactions (the absence/presence of a condition when a test falsely indicates that the condition is present/absent) and Psychiatric Status Rating Scale indicate that these articles

are grouped because of their involvement in extra-clinical research such as epidemiology/nosology, service development, and testing.

MedSOM echoes the organizational structure of the *KSCTP* textbook, albeit at a much higher level of abstraction. MedSOM has a "Neuropsychiatry" cluster, similar to the "Neural Sciences" section which opens all editions of the textbook. There are clusters specific to "Child psychiatry", and to "General/adult psychiatry". Psychiatry of old age is not separately represented and appears to be subsumed partly within the "General/adult psychiatry" cluster and partly within the "Neuropsychiatry" cluster. The "Social psychiatry" and "Psychoanalytic psychotherapy" clusters match up with chapters on "Contributions of the social sciences" and "Contributions of the psychological sciences". The "Administrative psychiatry" cluster incorporates elements from multiple chapters on the more systemic and technical features of psychiatry including "Quantitative and Experimental Methods in Psychiatry", "Theories of personality and psychopathology", "Diagnosis and psychiatry", "Classification in psychiatry", and "Public psychiatry".

The major structural difference between MedSOM and the textbook is that MedSOM does not appear to include any structural information related to diagnostic categories at either the individual diagnostic level or at the level of syndromes. For example, MedSOM does not include discrete sections for psychotic illnesses (such as schizophrenia) versus affective illness (such as depression and anxiety). The textbook relies heavily on diagnostic categories for its chapter structure. This difference appears to be partly due to the relatively small subset of information contained within the textbook reference sets compared with the MedSOM set (~20,000 textbook references versus 33 million used for training the MedSOM, including ~4 million addressing psychiatric topics).[24] Much of the knowledge related to diagnostic systems and differential diagnoses is contained within books such as the Diagnostic and Statistical Manual of Mental Disorders[31] which are not indexed by Medline. The focus of textbook references generally concerns cutting edge treatments and review articles which do not differentiate between individual diagnoses.

7.6.1 - Future research – optimizing the model by considering time and entropy

MedSOM is the first iteration of an approach to mapping the entire domain of peer-reviewed medical literature to make currently implicit knowledge explicit and observable by visual cues in two-dimensional maps. While it is encouraging that this form of the model provides a coherent and consistent interpretation of the knowledge contained within the *KSCTP* textbook, the model will need to be refined before it is ready to be integrated into the curriculum development process. In our opinion the two potential refinements most likely to improve the utility of the model are the addition of the capacity to consider the effects of time on the map of knowledge; and the introduction of a formal measure of the extent to which the SOM provides a coherent and consistent organizing structure for medical knowledge, such as entropy.

To understand how a temporal dimension may improve the MedSOM model, consider that one reason the MedSOM's structure does not include individual diagnoses as structural features may be the rapid changes in psychiatric nosology over the 50 years covered by the 10 editions of the textbook. In that time there have been 4 editions of the Diagnostic and Statistical Manual of Mental Disorders with multiple textual revisions, and with major methodological changes, particularly with the 3rd edition, which imposed a much greater focus on reliable diagnosis based on a phenomenological approach;[32] and the 5th edition, which moved away from categorical diagnoses towards a spectrum approach.[31] The MeSH used to identify diagnostic categories over this period have shown similarly rapid changes, which may have affected the ability of the MedSOM to extract stable patterns of knowledge, particularly given that the MedSOM was trained on data including articles dating back to the 19th century.

Nevertheless, the current research successfully demonstrates that a SOM trained on the 33 million articles indexed in the Medline database reveals implicit relationships between the subset of articles referenced by a core psychiatric textbook at a high level of abstraction. As the lack of temporal information in the MedSOM appears to have prevented a finer level of detail, extending the current

approach to a temporal SOM is a logical next step. The relative density SOM (ReDSOM) approach of Denny et al develops a longitudinal approach in which a series of SOM are trained using subsets of data divided into time periods, where the trained SOM of one period is used as the initial state of the SOM for the next period.[15]

The advantage of this approach is that where knowledge structures persist over consecutive time periods, they will continue to be reflected in the input data and therefore will persist in the SOM. Where knowledge structures change over time, the SOM will learn to replace the old structures with the new. For example, where a diagnosis has changed between time periods, reflected in a new MeSH, the ReDSOM can learn to associate the new MeSH with the existing clinical structures. We intend to extend the current approach to a temporal SOM using the ReDSOM approach. The ability to identify temporal changes in the organizational structure and relationships of the medical and psychiatric knowledge base would provide an empirical trigger for considering the introduction/removal of content into/out of medical and psychiatric curricula based on increasing/declining importance in the literature.

One of the most useful formal properties of network models like SOMs is their entropy, which is a measure of the disorder in a system. The MedSOM before training would be expected to have high entropy and be highly disordered, because each article would be expected to be equally likely to match any individual node. After training, articles with particular MeSH would be expected to be more likely to match nodes in particular regions, increasing the order and decreasing the entropy of the system.[33] Rousseau et al. (2019) discuss how the introduction of entropy to network models can be used to describe the integration of knowledge from multiple disciplines in the novel information represented by individual articles and regions of a SOM.[33]

7.6.2 - Limitations

The MedSOM reported in this research extracts relationships from data in an unsupervised manner, so that the meaning of its results have to be inferred. The current approach attempts to provide an

external validation of the meaning of the MedSOM with reference to a pre-existing knowledge structure created by experts. While this can strengthen confidence that the MedSOM has imputed structural relationships from the entire set of Medline articles that also apply to the subset of articles referenced by *KSCTP*, it is not directly verifiable. This involves a number of potential limitations. As with the current research, the meaning of the MedSOM may be concentrated at a very high level of abstraction. As a single step in a process of gradually more refined understanding, this is not necessarily a severe limitation, and in the current case, the addition of a temporal dimension has been suggested as a promising way to investigate more detailed levels of meaning in MedSOM and *KSCTP*.

7.7 - Conclusions

While maps of scientific and medical literature have great promise as empirical bases for endeavors such as medical curriculum development, their use is constrained by their limited intelligibility to domain experts. The current research demonstrates that an unsupervised SOM derived from the 33 million articles indexed by the Medline database can extract an organizational structure that coherently organizes the knowledge contained within a core psychiatric textbook, at a high level of abstraction. A key limitation of the current research is the use of a SOM trained on a set of articles published over a period of more than one hundred years. The extension of the current approach using a temporal SOM such as Denny's ReDSOM may allow for a more detailed organization of the knowledge indexed by the Medline database that more closely models the expert-defined structures of the *KSCTP* textbook.

7.8 - References

1. Harden RM. AMEE Guide No. 21: Curriculum mapping: A tool for transparent and authentic teaching and learning. Med Teach. 2001;23(2):123–37.

2. Thomas P, Kern DE, Hughes MT, Chen BY, editors. Curriculum Development for Medical Education: A Six-Step Approach. Third. Baltimore: Johns Hopkins University Press; 2015.

3. The Lancet. Cardiology's problem women. The Lancet [Internet]. 2019;393(10175):959. Available from: http://dx.doi.org/10.1016/S0140-6736(19)30510-0

4. National Library of Medicine. MEDLINE Database Home [Internet]. MEDLINE Home. 2021 [cited 2023 May 2]. Available from: https://www.nlm.nih.gov/medline/index.html

5. Skupin A, Biberstine JR, Börner K. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS One. 2013;8(3).

6. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Vol. 267, Artificial Intelligence. Elsevier B.V.; 2019. p. 1–38.

7. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. Applied Sciences (Switzerland). 2021 Jun 1;11(11).

8. Boyack KW, Klavans R. Creation and Analysis of Large-Scale Bibliometric Networks. In: Glänzel W, Moed HF, Schmoch U, Thelwall M, editors. Springer Handbook of Science and Technology Indicators. Cham, Switzerland: Springer Nature; 2019. p. 187–212

9. Kohonen T. Self-organizing maps. 3rd ed. Berlin: Springer; 2001.

10. Clarivate. Web of Science Home [Internet]. Web of Science Website. 2023 [cited 2023 May 2]. Available from: https://www.webofscience.com/wos/

11. Elsevier. Scopus Home [Internet]. Scopus Website. 2023 [cited 2023 May 2]. Available from: https://www.elsevier.com/en-gb/solutions/scopus

12. Klavans R, Boyack KW. Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge? J Assoc Inf Sci Technol. 2017 Apr 1;68(4):984–98.

13. Boyack KW, Newman D, Duhon RJ, Klavans R, Patek M, Biberstine JR, et al. Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. PLoS One. 2011;6(3):e18029.

14. Amos A, Lee K, Sen Gupta T, Malau-Aduli B. Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps. In: 19th World Congress on Medical and Health Informatics [Internet]. Sydney: Australasian Institute of Digital Health; 2023. Available from: https://raw.githubusercontent.com/AndrewAmosJCU/PsychSOM/main/ColorCoded.png

15. Denny, Williams GJ, Christen P. Visualizing temporal cluster changes using Relative Density Self-Organizing Maps. Knowl Inf Syst. 2010;25(2):281–302.

16. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli B. Identifying emerging topics in the psychiatric literature to facilitate curriculum renewal and development. Current Psychology. 2022 Oct

17. Silva A, Schrum M, Hedlund-Botti E, Gopalan N, Gombolay M. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. Int J Hum Comput Interact. 2022;

18. Shulner-Tal A, Kuflik T, Kliger D. Enhancing Fairness Perception–Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople's Fairness Perceptions of Algorithmic Decisions. Int J Hum Comput Interact. 2022;

19. He X, Hong Y, Zheng X, Zhang Y. What Are the Users' Needs? Design of a User-Centered Explainable Artificial Intelligence Diagnostic System. Int J Hum Comput Interact. 2022;

20. English JT. Review: Kaplan and Sadock's Comprehensive Textbook of Psychiatry, 7th ed., vols 1, 2. American Journal of Psychiatry. 2002 Feb;159(2):327–327.

21. Sadock BJ, Sadock VA, Ruiz P, editors. Kaplan and Sadock's Comprehensive Textbook of Psychiatry. 10th ed. New York: Wolters Kluwer; 2017.

22. Amos AJ, Sen Gupta T, Lee K, Malau-Aduli B. Mapping the evolution of psychiatric knowledge indexed in Medline 1900-2022 using self-organising maps. Knowl Inf Syst. 2023;In Preparation.

23. Ohniwa RL, Hibino A. Generating process of emerging topics in the life sciences. Scientometrics [Internet]. 2019;121(3):1549–61. Available from: https://doi.org/10.1007/s11192-019-03248-z

24. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Identifying emerging topics in the peer-reviewed literature to facilitate curriculum renewal and development. Current Psychology [Internet]. 2022 Dec 12; Available from: https://link.springer.com/10.1007/s12144-022-04090-y

25. Melka J, Mariage J. Efficient Implementation of Self-Organizing Map for Sparse Input Data. In: Proceedings of the 9th International Joint Conference on Computational Intelligence. Funcha, Madeira, Portugal; 2017. p. 54–63.

26. Melka J, Mariage JJ. Adapting Self-Organizing Map Algorithm to Sparse Data. In: 9th IJCCI 2017: Funchal, Madeira, Portugal (Selected papers). 2017. p. 139–61.

27. Kaplan HI, Sadock BJ, editors. Kaplan and Sadock's Comprehensive Textbook of Psychiatry. 4th ed. Baltimore: Williams and Wilkins; 1985.

28. Kaplan HI, Freedman AM, Sadock BJ, editors. Kaplan and Sadock's Comprehensive Textbook of Psychiatry. 3rd ed. Baltimore, MD: Williams & Wilkins; 1980.

29. National Library of Medicine. MEDLINE® Citation Counts by Year of Publication [Internet]. NLM Website. 2023 [cited 2023 Jun 25]. Available from: https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html

30. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020 Jun 1;58:82–115.

31. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th ed. Washington, DC: American Psychiatric Association; 2013.

32. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 3rd ed. Washington, DC: American Psychiatric Association; 1980.

33. Rousseau R, Zhang Lin, Hu X. Knowledge Integration: Its Meaning and Measurement. In: Glänzel W, Moed HF, Schmoch U, Thelwall M, editors. Springer Handbook of Science and Technology Indicators. Cham, Switzerland: Springer Nature; 2019. p. 73-94.

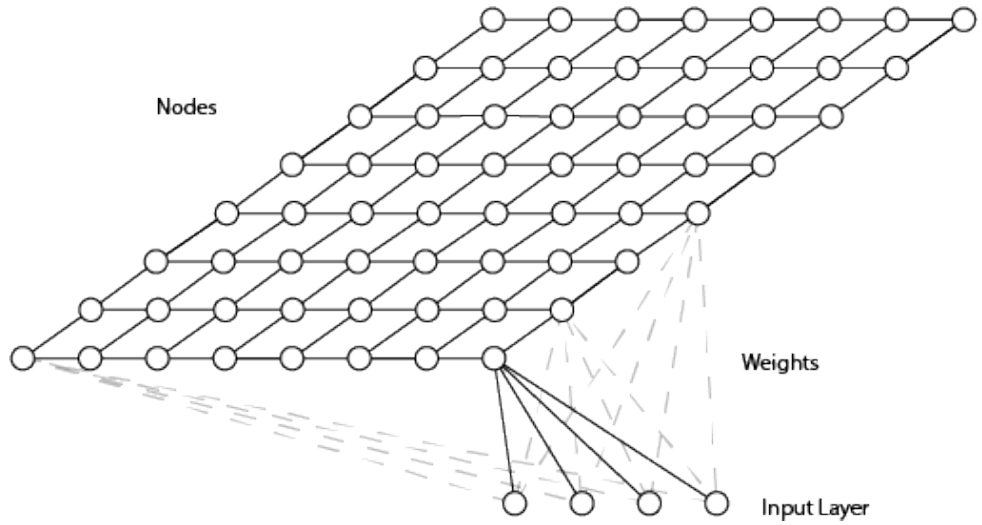
7.9 - Supplementary Material 1: Technical features of self-organizing maps

Self-organizing maps (SOMs) are a type of neural network composed of a grid of nodes each of which is an independent unit that processes inputs from the environment and which learn over time to recognize specific features of those inputs. To understand SOMs it is necessary to understand: 1) what a SOM does; 2) the representation of information in the SOM; 3) how individual nodes process inputs from the environment; 4) how the whole SOM learns from inputs; and 5) how the SOM represents what it has learned.

<u>What a SOM does and why this is useful</u>

- A SOM transforms high-dimensional information into a 2-dimensional map which retains many of the properties of the high-dimensional form; In Supplementary Figure 7-1 the four dimensions of the input layer are transformed into the 2 dimensions of the grid of nodes
- In particular, a SOM will retain the property that inputs that were close together in the high-dimensional space will be close together in the 2-dimensional space (topological ordering)

Supplementary Figure 7-1: Self-organizing map (8x8 node grid; 4-dimensional input layer; 4 weights per node):



- This is useful because human vision is highly effective and efficient at understanding spatial relationships in 2-dimensions and 3-dimensions, but completely unable to understand higher-dimensional relationships

<u>How information is represented in self-organizing maps</u>

- SOMs represent information as vectors, which are ordered sets of numbers of particular size where the position has meaning; in Supplementary Figure 7-1 the input layer is depicted as a vector of four numbers, and each node receives four weighted inputs from the input layer
- The vectors in this research project are sets of 29,917 numbers where the position represents a specific Medical Subject Heading (MeSH)
- In computer memory, this looks like a set of 1s and 0s: [0, 1, 0, 0…..(x29,917)]
- A vector with a 1/0 in a particular position means the presence/absence of a particular MeSH, for example the vector in the above line might mean: [“Schizophrenia” absent/0, “Bipolar disorder” present/1, “ADHD” absent/0, “Anxiety” absent/0, ….etc]

Representation of articles: The representation of an article is the pattern of 1s and 0s in a vector representing all the MeSH annotating that article (as each article is annotated by an average of 9 MeSH, most of the positions will be 0s)

Vector representing an article (7 numbers shown, 29,910 numbers represented by …):

0	1	0	0	0	…	1	0

Representation of nodes: Each node represents a neuron in the brain, which receives information, processes the information, and produces a level of activation. Each node is specified as a vector of the same length as the inputs (29,917 in this research) where each position in the vector represents a weight which is used to process inputs. Weights are not 0s and 1s, but floating-point numbers which vary between certain limits (e.g., between 0.0 and 1.0)

Vector representing a node (7 numbers shown, 29,910 numbers represented by …):

0.11	0.76	0.01	0.003	0.01	…	0.98	0.001

<u>How individual nodes process inputs from the environment</u>

- In this research, the inputs are articles represented by vectors which specify the MeSH which annotate those articles (as above)
- When an article is presented to the SOM, every node independently processes the information from the vector representing that article
- Each node calculates the distance squared between the weight of the node at a particular position, and the article value at that position

Article

1	0	0	1	0	…	1	0

Vector -

0.11	0.76	0.01	0.003	0.01	…	0.98	0.001

Distance =

0.89 x 0.89	(-0.24) x (-0.24)	(-0.01) x (-0.01)	0.997 x 0.997	(-0.01) x (-0.01)	...	0.02 x 0.02	(-0.001) x (-0.001)

SUM SQUARES = 1.84 [ignoring the other 29,910 MeSH positions]

- The node with the lowest sum of squares after input from a particular article is described as the best-matching unit or BMU; this node is the one where the weight vector is most like the article vector

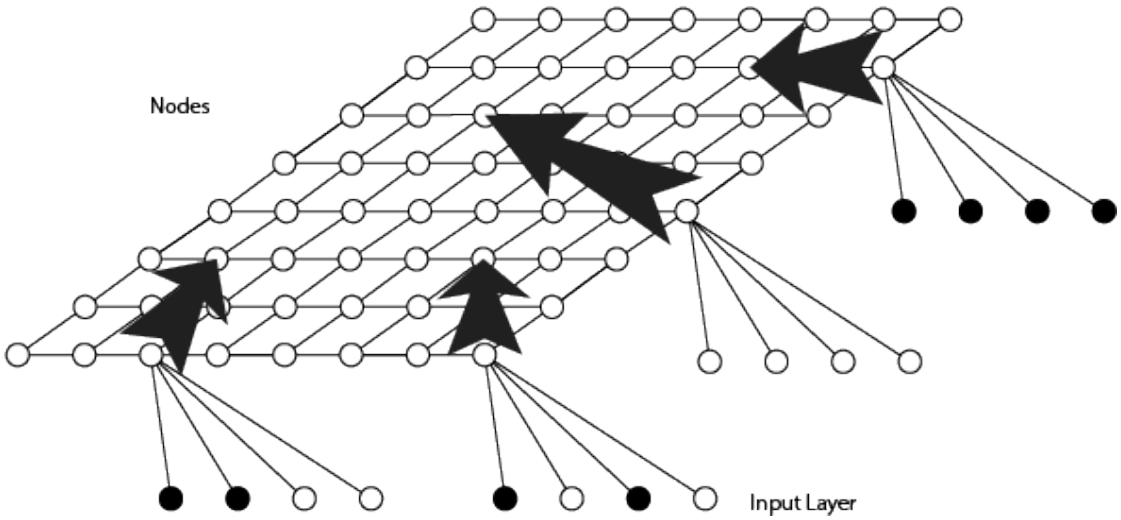
How the whole SOM learns from inputs

- In the simplest case, before training starts the weights of every node in a SOM will be set to random numbers
- As a result, the BMU of each article will be more or less randomly distributed across the grid of nodes, because all article vectors will be equally likely to be similar to all node vectors
- The SOM learns by changing the node weights of the BMU after each presentation of an article by moving those weights closer to the article vector; it also moves the weights of the nodes immediately surrounding the BMU towards the article vector, but moves them less than the weights of the BMU
- Over multiple iterations of presentation of all articles to the SOM, this process leads to topological ordering of the SOM grid such that inputs that were close together in the high-dimensional space will be close together in the 2-dimensional space (topological ordering; Supplementary Figure 7-2)

Supplementary Figure 7-2 shows four inputs with different vectors learning to differentiate themselves over multiple iterations by activating spatially diverse nodes separated on the face of the grid. Black dots represent true/1 and white dots represent false/0 in the inputs. Over learning the BMU shifts from the original random node to a node which reproduces closeness in the high-

dimensional space represented by the input (4-dimensions in Supplementary Figure 7-2) by closeness in the 2-dimensional grid of the SOM.

Supplementary Figure 7-2: Learning in Self-organizing map:



7.10 - Supplementary Material 2 – Identifying implicit bias using self-organizing maps

One of the most powerful features of SOMs is their capacity to superimpose visualizations derived from different dimensions of the high-dimensional conceptual space on which they are trained. This can be used as a form of high-dimensional Venn diagram that allows for the isolation of areas of knowledge that share specific characteristics in ways useful for curriculum development, for example by identifying areas of research that are both primarily psychiatric, and include significant differences between the sexes.

To illustrate how this could be used for the purposes of reducing the sorts of medical research biases regarding gender noted in the main text, it is useful to consider the existing approach of Sex and Gender Specific Health (SGSH), which "aims to understand sex- and gender-based differences in diseases common to both women and men, with the goal of applying the sex and gender-specific knowledge into clinical practice to improve patient outcomes" (p181).[1]

As a tool for detecting and correcting the history of ignoring differences between the sexes due to practices such as excluding women from medical trials due to the potential for pregnancy, Song et

al. (2016) created a SGSH searchable database which identified the subset of Medline-indexed articles likely to contain SGSH data using a combination of specific MeSH. Individual searches could then be applied to identify SGSH articles on specific diseases such as diabetes.[1]

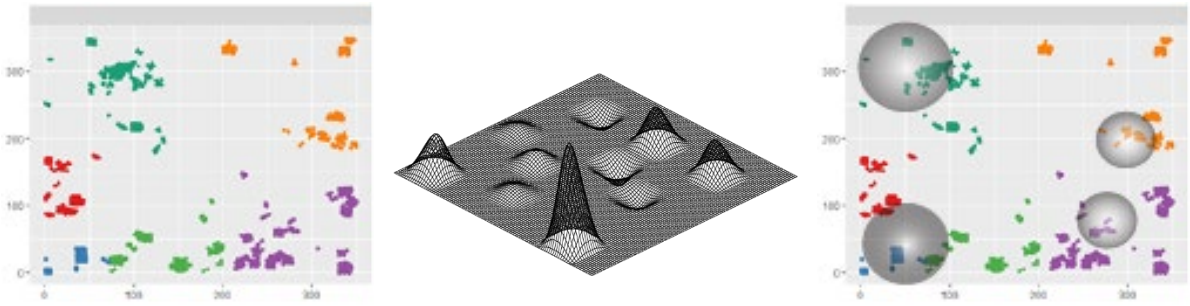
By mapping the high-dimensional space representing all MeSH to a 2-d surface, the SOM allows for the simultaneous application of this technique to all medical knowledge simultaneously. It is possible to represent the degree of importance of subsets of MeSH, such as those used to generate the SGSH database above, across all nodes of a SOM by changing the weights of all other MeSH to 0, and representing the results in a density map (see Supplementary Figure 7-3).

Supplementary Figure 7-3: Superimposing the SGSH density map on the MedSOM

a) Textbook SOM projection

b) SGSH density map

c) Density x SOM projection



Note: Supplementary Figures 7-3a) and 7-3c) are taken from the main analysis of the paper, but the SGSH density map is purely illustrative; this analysis has not been done.

Supplementary Figure 7-3 shows how the SOM makes it possible to superimpose the SGSH density map (b) representing the prominence of research featuring SGSH data across all the articles indexed by the Medline database, on the projection of the subset of knowledge contained within the KSCTP textbook (a). By considering only the peaks in the density map, which indicate the areas of knowledge with the most SGSH data, it becomes possible to identify the parts of the KSCTP addressed in the textbook which could start to consider differentiating sex and gender as part of the curriculum (c).

This method of identifying existing information about which curriculum developers were previously unaware is potentially useful for reducing implicit biases caused by a lack of knowledge. At a systemic level, superimposing the troughs of the SGSH density map on the SOM could also identify the areas of medical (or specifically psychiatric) knowledge with the least SGSH data. This would be expected for sexually dimorphic categories of disease, like prostate cancer or ovarian cancer, but it might indicate gaps in knowledge for diseases strongly affecting both sexes, such as heart disease.

References

1. Song MM, Simonsen CK, Wilson JD, Jenkins MR. Development of a PubMed Based Search Tool for Identifying Sex and Gender Specific Health Literature. J Women's Health. 2016;25(2):181-187

Chapter 8: Testing the validity of emerging topics in psychiatry identified by a machine learning algorithm with a sample of working psychiatrists

Authors:

Andrew James Amos[*1], MB.BS, Kyungmi Lee[2], PhD, Tarun Sen Gupta[1], PhD, Bunmi S. Malau-Aduli[1,3], PhD

* Corresponding Author

1 College of Medicine and Dentistry, James Cook University, Townsville, Australia

2 College of Science and Engineering, James Cook University, Cairns, Australia

3 School of Medicine and Public Health, University of Newcastle, Newcastle, Australia

This chapter describes an experiment intended to test the validity of the Increment statistic as an indicator of the novelty and relevance of Medical Subject Headings used to annotate articles in the Medline database of peer-reviewed medical literature. It asked practising Fellows of the Royal Australian and New Zealand College of Psychiatrists to rank articles identified by the Increment statistic in order of their novelty and separately in order of their relevance to practice. It compared these rankings with articles selected at random and found that the Increment statistic did identify articles ranked as significantly more novel and relevant than randomly selected articles.

8.1 - Abstract

Background: This study investigated the validity of the output of a machine learning algorithm designed to identify emerging topics in the medical literature by asking practicing psychiatrists to rank the novelty and relevance of scientific articles chosen by the algorithm.

Methods: Twenty-one practicing psychiatrists completed an online questionnaire by ranking the novelty and relevance of 5 articles selected by a machine-learning algorithm and 5 randomly selected psychiatric articles.

Results: The psychiatrists ranked the machine-learning selected articles as significantly more novel ($p<.03$) and significantly more relevant ($p<.01$) than the randomly selected articles. The novelty and relevance of individual articles were not closely correlated ($\rho = -0.18$, 95% CI (-0.32, -0.02).

Conclusions: The validity of the machine-learning algorithm for detecting emerging topics in the peer-reviewed medical literature was supported by the fact that practicing psychiatrists ranked articles selected by the algorithm as more novel and relevant than randomly selected articles. The utility of the machine-learning algorithm can be greatly increased by integrating information about individual clinicians, such as their area of practice and the peer-reviewed articles they have recently read.

Keywords:

artificial intelligence; machine learning; medical education; bibliometrics; curriculum development

8.2 - Background

The practice of medicine depends on an enormous, diverse, and continually increasing set of scientific evidence, clinical knowledge, and administrative frameworks.[1] The rapid development of machine learning (ML) and artificial intelligence (AI) techniques capable of extracting useful information from large datasets, including patterns not otherwise intelligible to individual humans,

has tremendous potential for streamlining the learning processes involved in medical training and practice from medical school all the way through to retirement.[2]

One of the largest barriers to the use of ML in medical education and training is that it can be difficult for humans to understand how ML processes work and what they mean.[3] Perhaps due to their fiduciary responsibilities for patients and other stakeholders, doctors are even more reluctant to rely upon opaque ML processes than other experts.[4–6] One way to reduce this resistance is to validate the results of ML processes against meaningful empirical data.

We recently reported on a technique that validated an artificial intelligence derived map of all the medical knowledge indexed by the Medline database of peer-reviewed medical literature. This approach examined whether the structure and organisation automatically extracted from Medline by a self-organising map consistently and coherently organised the references of a psychiatric textbook.[7]

Having confirmed that an AI-derived map was consistent with the organisational structure of a textbook, we wanted to develop the capacity to validate ML outputs directly with human subjects. The current research attempts to validate the results of a ML algorithm that analyses psychiatric research by testing whether it generates information that is meaningful to a group of practicing psychiatrists.

8.2.1 - Continuing professional development within Messick's validity framework

The articles selected by a ML as likely to be both novel and relevant to a group of practicing psychiatrists can be conceptualised within Messick's validity framework. It is widely recognized that the knowledge, skills, and abilities that medical doctors develop over the course of their university education and specialist training must be kept current by continuing professional development (CPD)/continuing medical education (CME) throughout their careers. In recognition of the autonomy

and discretion expected of doctors, CPD programs require a minimum amount of time spent on various types of CPD activities but leave the content up to doctors.[8]

The requirement that doctors annually demonstrate an adequate level of CPD can be likened to the milestones in general medical education discussed by Hamstra & Yamazaki.[9] Their argument suggests that because the health professions are largely self-regulated, the system of CPD can be considered analogous to a complex assessment system capable of measuring the extent to which individual doctors have maintained professional standards of knowledge, skills, and attributes (KSA).

Hamstra and Yamazaki[9] summarised Messick's framework in 5 essential elements of validity: 1) Content is the extent to which items are representative of the construct being measured (for CPD, the maintenance of sufficient KSA for competent professional practice); 2) Response process is the evidence of data integrity comprising clear instructions and reliable ratings; 3) Internal structures are psychometric properties including reliability, difficulty, and the latent dimensions that comprise the overall construct; 4) Relations with other variables including convergent and divergent evidence; and 5) Consequences, particularly impact on learners and the system in which they operate.

Hamstra and Yamazaki[9] conclude that a useful way of addressing the complexity of learning programs like CPD is to analyse "the validity of the decisions drawn from the data to allow for more effective tools for program directors…to…improve their training programs and clinical learning environments."[9] (p78)

Swiecki et al.[10] examine the promise and pitfalls of AI techniques using Messick's three models of assessment design. The *student model* describes the KSA to be acquired; the *task model* describes what students do to generate evidence they have acquired these KSA; and the *evidence model* describes how tasks can be used to make inferences about student characteristics.

Combining these frameworks suggests it is possible to evaluate the validity of AI products for CPD/CME by examining how they might improve training programs and/or learning environments by their impact on one or more of the student, task, and evidence models. Swiecki et al.[10] summarise the possibilities:

> "In terms of the student model, the presence of AI suggests that we should adjust the traits, skills, and abilities assessed to be those that require human influence rather than those that AI can accomplish on their own. In terms of the task model, AI suggests that we should allow students to use AI-based computational tools during the assessment. And in terms of the evidence model, the presence of AI suggests that we should account for the fact that a human-AI team can generate assessment evidence." (p8)

It seems reasonable to assume that medical consultants can efficiently and effectively prioritize the areas they need to focus on to maintain broad-ranging skills and knowledge and develop specific areas of interest. However, the volume and complexity of new medical knowledge makes it difficult for any individual doctor to be aware of the entire range of topics that might improve their practice.[8] There is an opportunity for bibliometric and ML techniques that rapidly evaluate, filter, and summarize enormous sets of data to help doctors optimize their CPD activities by focusing their attention on information that is both intellectually novel and relevant to practice.[2] This may have the additional advantage of identifying and/or reducing the biases that can arise when medical curricula are based entirely on expert judgement.[7]

8.2.2 - Machine learning and continuing professional development

The most common automated method of establishing the importance of published peer-reviewed research is to analyse citation patterns. This type of analysis reveals a characteristic pattern in the emergence of important new research, where a large number of articles cited a small number of

times all reference a small set of seminal papers each of which is cited a large number of times.[11] Innovation can be detected when the seminal papers change.

While this is an effective method of evaluating the impact of individual articles, clear citation patterns at the level of individual articles take years to emerge.[11] Ohniwa et al. developed a more immediate way to identify topics of emerging interest in the peer reviewed medical literature.[12,13] They used the Medline database, which indexes the set of articles published in the most reliable peer-review medical journals, with records dating back to the 19th Century, to develop an Increment statistic that ranks medical subjects in order of increasing research activity. [12,13]

The US-based National Library of Medicine curates the Medline database, and it maintains a controlled vocabulary of all the medical topics addressed across all the indexed articles, which it calls Medical Subject Headings (MeSH). Each indexed article is tagged with multiple MeSH which summarise the medical topics they address. The Increment statistic developed by Ohniwa et al.[12,13] analysed the frequency with which articles indexed in the Medline database were tagged with particular Medical Subject Headings (MeSH) compared with prior and following years. They reasoned that MeSH which were suddenly being used much more in peer reviewed articles were likely to be good indicators of topics of emerging interest in medical research, independent of the absolute number of tags.

In previous research, we used the Increment statistic to identify emerging topics within the subset of MeSH devoted to psychiatric and psychological subjects.[12] In this research, we used an online questionnaire to test whether scientific articles which addressed emerging topics were perceived as more novel and relevant to the practice of a group of practicing psychiatrists than randomly selected articles which did not address emerging topics.

With respect to Messick's model of validity as an argument from evidence, confirming that a ML algorithm is capable of predicting how novel and relevant a set of articles is to a group of practicing

psychiatrists has applications for all three models, particularly as a demonstration of the principle that human-AI interaction can generate assessment evidence.

We hypothesised that articles indexed in the PubMed database that were labelled with MeSH identified as representing emerging topics in a particular year (2022) would be ranked more highly than articles selected at random from the set not labelled with emerging topic MeSH.

8.3 - Methods

8.3.1 - Creation of questionnaire

The questionnaire (Supplementary Material 1) was designed to include two ranking exercises, in which participants would be asked to rank a single list of article titles twice, once in order of the novelty of their content to the participant, and once in order of their relevance to participants' practice. After experimenting with different frameworks and considering statistical power it was decided that the list of article titles would be ten items long, comprising five articles selected as representatives of emerging topics, and five selected at random.

The set of article titles was sampled from the larger set of articles indexed in PubMed, published in year 2020, and where the article was annotated with at least one MeSH indicating a psychiatric/psychological topic. This set of articles was further divided into two mutually exclusive groups, with the emerging topic group containing all articles annotated with at least one of the top 10 emerging topic MeSH for the year 2020 (see Table 8-1); and the comparison group containing all articles not annotated by at least one of those 10 emerging topic MeSH.

Table 8-1: Top 10 Emerging Topic Mesh for year 2020

Code	MeSH
F03.900.647.500	Opioid related disorders
F01.145.126.990.367.500	Burnout, professional

F02.830.900.333.500	
F03.625.164.113	Autism spectrum disorder
F02.940	Resilience, psychological
F01.145.126.980.875.149	Suicidal ideation
F01.145.813.840	Social Stigma
F01.752.723	Neuroticism
F04.711.513	Neuropsychological tests
F03.615.250.700	Cognitive dysfunction
F03.615.400.100	Alzheimer disease

In order to generate the final list of ten items, five articles were selected at random from the list of articles labelled with at least one emerging topic psychiatric/psychological MeSH; and five articles were selected at random from the list of articles with no emerging topic psychiatric/psychological MeSH annotations (Table 8-2). Random article selection was achieved using the R programming language version 4.3.0 function *sample* to select 5 locations in a vector listing the unique Pubmed ID of every article in each category.

Table 8-2: List of selected article titles

Emerging topic articles			
Article Title	Emerging topic MeSH annotating article	Novelty (mean, min/max)	Relevance (mean, min/max)
Features of personality and professional	Burnout, professional Neuroticism	6.5 1 - 10	4.8 1 - 10

burnout syndrome of physicians - analysis based on questionnaires studies.			
Using Natural Language Processing and Machine Learning to Identify Hospitalized Patients with Opioid Use Disorder	Opioid related disorders	4.6 1 - 10	4.8 1 - 9
Identifying Clinical Risk Factors for Opioid Use Disorder using a Distributed Algorithm to Combine Real-World Data from a Large Clinical Data Research Network	Opioid related disorders	5.7 1 - 10	4.3 1 - 8
Mental Health Comorbidity Analysis in Pediatric Patients with Autism Spectrum Disorder Using Rhode Island Medical Claims Data	Autism spectrum disorder Suicidal ideation	5.9 2 - 10	4.8 1 - 10

Relationship of Neurofilament Light (NfL) and Cognitive Performance in a Sample of Mexican Americans with Normal Cognition, Mild Cognitive Impairment and Dementia	Alzheimer disease Cognitive dysfunction Neuropsychological tests	2.9 1 - 10	6.2 1 - 10
Comparison articles			
Reactions on 'The effectiveness of monoamine oxidase inhibitors in treatment-resistant depressive disorders in clinical practice; a retrospective open-label study'	None	6.6 1 - 10	5.0 1 - 10
Restricted Effect of Cerebral Microbleeds on Regional Magnetic Susceptibility	None	4.9 1 - 9	6.5 1 - 10
Patient preferences when searching for	None	5.8 2 - 10	4.8 1 - 8

clinical trials and adherence of study records to ClinicalTrials.gov guidance in key registry data fields.			
No need to touch this: Bimanual haptic slant adaptation does not require touch.	None	4.9 1 - 10	7.6 1 - 10
Motor functions: Motor development includes the evolution from reflexive to voluntary and goal-directed motor actions.	None	7.2 2 - 10	5.9 1 - 10

8.3.2 - Participants and Recruitment

The Royal Australian and New Zealand College of Psychiatrists (RANZCP) is the professional body in those countries with oversight of psychiatric training prior to Fellowship[14] and CPD afterwards.[15] Every month the RANZCP and each of the individual states of Australia and NZ sends an email newsletter to all psychiatric consultants on an opt-out basis, reaching more than 4000 psychiatrists across Australia and New Zealand. A section of these newsletters is designed to allow Fellows to communicate with their colleagues for purposes such as recruitment for research questionnaires. After designing our questionnaire, we recruited psychiatric consultants with a single paragraph invitation and link directly to the questionnaire in these newsletters. Informed consent was obtained

from all participants, and the research project was approved by the Human Research Ethics Committee of James Cook University (Ethics Approval Number: H9432). Use of the Psyche newsletter was approved by the RANZCP contingent upon the James Cook University ethics approval.

8.3.3 - Statistical Analysis

An a priori power analysis suggested 14 participants ranking 10 articles, where 5 were selected for their novelty and 5 were randomly selected, using Wilcoxon signed rank test with matched samples, would have adequate power (80%) for detecting a medium sized effect with a statistical significance criterion level of 5%. Wilcoxon rank sum tests were used to examine whether participants judged the ML selected and randomly selected article titles to be different with respect to novelty and relevance to practice.

8.4 - Results

In all, 21 participants completed both ranking tasks. Wilcoxon rank sum tests indicated that participants ranked the ML-selected article titles as significantly more novel and significantly more relevant than the randomly selected article titles (Table 8-3).

Table 8-3: Wilcoxon rank sum test of novelty and relevance of ranked articles

Condition	Median rank - Selected	Median rank - Random	W	sd	p
Novelty	5	6	4240	409.3	<.03
Relevance to practice	5	7	3974	409.3	<.01

Figure 8-1 and Tables 8-2 and 8-3 show that the five articles selected by the ML algorithm were ranked as both more novel and more relevant than five randomly sampled articles.

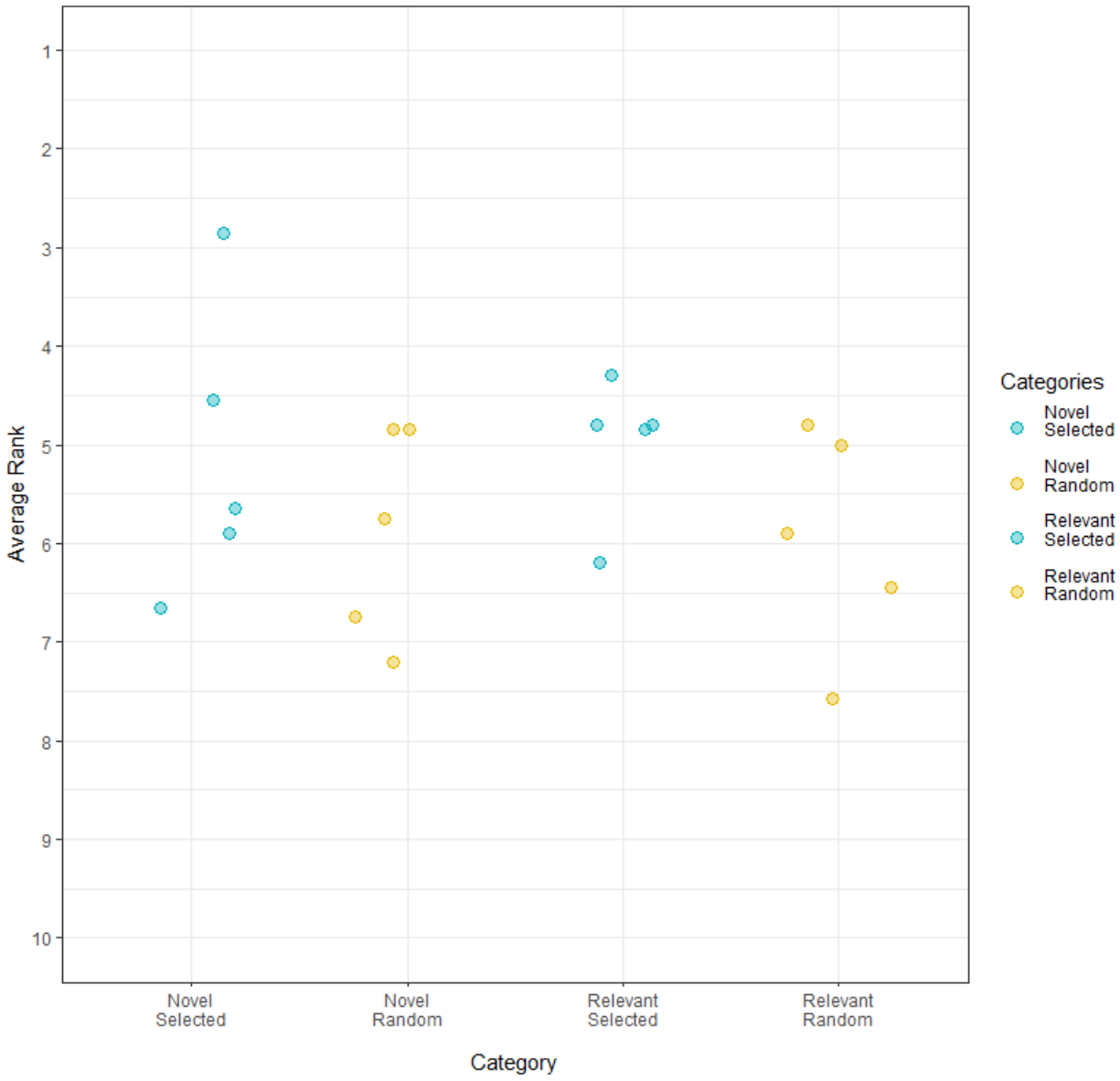


Figure 8-1: Rank distribution of articles selected by ML algorithm or selected randomly

Other relevant features of Figure 8-1 include that the best ranked article for novelty was among those selected by the ML algorithm (top of first column; see also Table 8-2). The same article was the worst ranked for relevance in the ML selected category (bottom of third column), although it was ranked as more relevant than two of the randomly selected articles (bottom of fourth column). This is consistent with title of the article, “Relationship of Neurofilament Light (NfL) and Cognitive

Performance in a Sample of Mexican Americans with Normal Cognition, Mild Cognitive Impairment and Dementia", which concerns an obscure genetic marker (NfL) that most psychiatrists would be unaware of. It makes sense that this would be both highly novel and lack relevance to practice.

Post-hoc, in response to the indication that novelty and relevance rankings were not highly correlated, we calculated Spearman's rank correlation coefficient. We found a ρ of -0.18 with 95% confidence intervals between -0.32 and -0.02.

8.5 - Discussion

This research showed that a sample of practicing psychiatrists ranked a set of articles, selected for characteristics identified by a ML algorithm as emerging topics in the fields of psychiatry and psychology, as significantly more novel and relevant than a set of randomly selected articles within those fields. This provides evidence for the validity of the information about the characteristics of peer reviewed research extracted by a ML algorithm which may be useful for guiding decisions about continuing professional development.

In Swiecki's formulation,[10] AI can contribute to understanding of the KSAs required to maintain professional competence (the *student model*); the activities that demonstrate that students have those KSAs (the *task model*); and how those activities can be used to make inferences about those KSAs (the *evidence model*). An ML algorithm capable of predicting which articles will be most novel and relevant to practising psychiatrists could contribute to all three models. The most promising use of the algorithm would be as part of a human-AI interaction that leveraged ML capacity to rapidly analyse an enormous body of literature and present a short-list to the expert who would leverage the human capacity to recognise meaning and value to select the material with the greatest learning yield.

However, the results also make it clear that utilising the information provided by ML and AI algorithms to support professional activities like CPD is likely to require iterative refinement before it

can be implemented in practice. Overall, there was a small negative correlation between novelty and relevance rankings, and the most highly ranked novel article identified by our ML algorithm was also ranked as one of the least relevant to practice. This is not entirely surprising, as by definition the most innovative research concerns new methods that have not yet been integrated into practice. This particular article also illustrates that emerging topics often involve the basic sciences that underlie medicine rather than their application to clinical practice.

Fortunately, there are many possible options to improve the relevance of articles selected by ML or AI algorithms. Our research used anonymous responses to an online questionnaire with a fixed set of articles, so we were unable to calibrate the article selection using individual characteristics.

The next step in improving the relevance of articles selected for their novelty would be to refine the set of emerging MeSH topics using information about individual clinicians. This information could be automatically generated from existing data, for example by matching an individual clinician's specialist designation to the relevant MeSH topics, or tracking recent articles read by each clinician and selecting emerging topic MeSH most similar to their recent interests; or it could be manually collected, for example by asking clinicians to nominate their topics of interest or select from a list of emerging topics.

In Swiecki et al.'s[10] and Hamstra's[9] formulations this type of human-AI interaction could contribute to all three models of assessment. It could efficiently and effectively identify new topics to expand a consultant's KSA (content validity within the *student model*); simultaneously provide evidence that the consultant had considered a broad range of literature and engaged in detailed study of novel and relevant articles (response processes within the *task model*); and link those articles to specific KSAs (internal structure within the *evidence model*).

A more tailored version of this ML algorithm could be anticipated to have significant consequences for learners and CPD/CME frameworks. The human-AI interaction could improve the quality of learning, reduce the time spent in administrative tasks (including finding relevant literature and

demonstrating its relevance), and facilitate its integration into broader learning objectives, for example by iteratively improving the algorithm's understanding of each individual clinician's interests.

Ultimately, the purpose of the current research was to examine the validity of the emerging topic Increment statistic in selecting MeSH for novelty and relevance. The fact that articles selected with Increment statistic identified MeSH were ranked higher than average for novelty and relevance supports the validity of this technique.

8.5.1 - Limitations

The main limitations of the study are a consequence of its exploratory nature. It is the first study that attempts to support the validity of a ML algorithm for the selection of novel and relevant articles by asking content experts to rank the results and so cannot be judged against similar work. The study is the first step in a series of refinements that would be required to make the results of the ML algorithm useful for individual clinicians by personalising the selection process using each individual's preferences, either manually or using automated processes. Finally, while the participants were all practicing psychiatrists in Australia and New Zealand, this is a heterogeneous group. It is likely that there would be systematic differences in the novelty and relevance of any set of articles for psychiatrists who differ by characteristics including: private/public practice; early career/late career; and area of specialist practice (for example pediatric versus adult versus geriatric psychiatry).

Methodological limitations include the small number of participants and questions, although formal analysis suggested our sample did have adequate power to find a significant result if one existed. As the first study of its kind our results may be subject to social desirability bias, although the small number of data points and anonymised responses prevent the identification of such bias using statistical methods.[16]

8.6 - Conclusions

The fact that a group of practicing psychiatrists ranked a set of articles chosen by a ML algorithm higher for novelty and relevance than randomly selected articles supports its validity. The effect size of the algorithm can be improved by personalising its selections using information about individual clinicians. This could involve both automatically generated information, for example by considering the features of the scientific articles recently read by each clinician; and manually collected information, for example by asking for a list of current interests.

The final step in utilising ML/AI algorithms to improve CPD for medical practitioners will always involve practitioners' judgement in evaluating the worth of the recommendations made. While a ML algorithm like that described here can provide a list of articles that is likely to be novel to the practitioner and relevant to their practice, it will always be up to them to decide how valuable it will be to maintaining or expanding their expertise.

8.7 - List of abbreviations

AI	Artificial intelligence
CME	Continuing medical education
CPD	Continuing professional development
MeSH	Medical subject heading
ML	Machine Learning
NfL	Neurofilament light
RANZCP	Royal Australian and New Zealand College of Psychiatrists

8.8 - Declarations

Ethics approval and consent to participate: Consent to participate was sought from all participants as part of the online questionnaire, after the information sheet. Participants who did not provide consent were thanked and the questionnaire was not administered. Ethics approval was obtained from the Human Rights Ethics Committee of James Cook University Ethics Approval Number: H9432) on 07/05/2024.

Consent for publication: Consent to publish/present was sought from all participants as part of the online questionnaire, after the information sheet. Participants who did not provide consent were thanked, the questionnaire was not administered, and no data was collected.

Availability of data and material: All data and materials available on request. All data were collected anonymously.

Competing interests: The authors declare they have no competing interests

Funding: No funding was used for this project.

Author contributions: All authors contributed to the study conception and design in the context of providing support for the PhD of the corresponding author. Material preparation, data collection, and analysis were performed by Andrew James Amos. The first draft of the manuscript was written by Andrew James Amos and all authors collectively contributed to drafts of the manuscript.

Acknowledgements: Not applicable.

8.9 - References

1. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc [Internet]. 2011;122(319):48–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21686208%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3116346

2. Azer SA, Guerrero APS. The challenges imposed by artificial intelligence: are we ready in medical education? Vol. 23, BMC Medical Education. BioMed Central Ltd; 2023.

3. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020 Jun 1;58:82–115.

4. Miller T. Explanation in artificial intelligence: Insights from the social sciences. Vol. 267, Artificial Intelligence. Elsevier B.V.; 2019. p. 1–38.

5. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: A systematic review. Applied Sciences (Switzerland). 2021 Jun 1;11(11).

6. Silva A, Schrum M, Hedlund-Botti E, Gopalan N, Gombolay M. Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. Int J Hum Comput Interact. 2022;39(7):1390–404.

7. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Validating the knowledge represented by a self-organizing map with an expert-derived knowledge structure. BMC Med Educ. 2024;24(416).

8. Institute of Medicine. Best care at lower cost : the path to continuously learning health care in America. Washington, D.C.: National Academy of Sciences; 2012.

9. Hamstra SJ, Yamazaki K. A Validity Framework for Effective Analysis and Interpretation of Milestones Data. J Grad Med Educ. 2021 Apr 1;13(2s):75–80.

10. Swiecki Z, Khosravi H, Chen G, Martinez-Maldonado R, Lodge JM, Milligan S, et al. Assessment in the age of artificial intelligence. Computers and Education: Artificial Intelligence. 3(2022): 100075.

11. Glänzel W, Thijs B. Using hybrid methods and 'core documents' for the representation of clusters and topics: the astronomy dataset. Scientometrics. 2017; 111:1071-1087.

12. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Identifying emerging topics in the peer-reviewed literature to facilitate curriculum renewal and development. Current Psychology [Internet]. 2022 Dec 12;42:30813–24. Available from: https://link.springer.com/10.1007/s12144-022-04090-y

13. Ohniwa RL, Hibino A. Generating process of emerging topics in the life sciences. Scientometrics [Internet]. 2019;121(3):1549–61. Available from: https://doi.org/10.1007/s11192-019-03248-z

14. RANZCP. Training Program Overview [Internet]. RANZCP.org. 2024 [cited 2024 Sep 5]. Available from: https://www.ranzcp.org/training-exams-and-assessments/fellowship-program/program-overview

15. RANZCP. CPD Overview [Internet]. RANZCP.org. 2024 [cited 2024 Sep 5]. Available from: https://www.ranzcp.org/cpd-program-membership/cpd-program/cpd-overview

16. Leite WL, Cooper LA. Detecting social desirability bias using factor mixture models. Multivariate Behav Res. 2010 Mar;45(2):271–93.

8.10 - APPENDIX: Questionnaire

The questionnaire described in this article was hosted at:

https://jcu.syd1.qualtrics.com/jfe/form/SV_2bDqwNJvMNyXIkC

The information sheet, consent process, and sequence of questions is reproduced below as a series of image files. The originals were presented as browser pages.

Questionnaire Page 1 – Information sheet and consent to participate.

Information Sheet- Ranking the novelty and professional educational value of psychiatric research identified by data mining peer reviewed scientific literature

Sponsor: James Cook University, College of Medicine and Dentistry. This study has been approved by the Human Research Ethics Committee (HREC Approval Number: H9432).

Principal Investigator: Dr Andrew Amos, F.RANZCP, Ph.D Candidate - James Cook University. andrew.amos@jcu.edu.au

Eligibility:

We are seeking to recruit Fellows of the RANZCP who are currently required to participate in Continuing Professional Development (CPD) in order to maintain their registration as psychiatrists.

What is the purpose of the research?

The study is being conducted by Ph.D candidate Dr. Andrew Amos and the results will contribute to a degree in Doctor of Philosophy (Health) at James Cook University. The research is part of project that applies Artificial Intelligence techniques to the peer reviewed medical literature in order to generate an empirical basis for medical curriculum development, using psychiatric knowledge as a test case. A previous study by the same authors analysed the complete set of English-language articles indexed in the PubMed/Medline database to identify "emerging topics" of interest based on research activity.[1] Related research generated a map of the knowledge contained within PubMed/Medline distinguishing between general medical and psychiatric domains.[2]

This research seeks to validate the ability of the automated AI processes to identify research that is novel and relevant to psychiatrists in order to provide an unbiased empirical basis for selecting content for inclusion in CPD and curriculum development.

What are participants being asked to do?

The previous research has been used to select 10 articles from peer-reviewed psychiatric literature published in 2022, with 5 being selected using the AI generated information as more novel and 5 being selected at random. Participants are asked to read the titles and abstracts of the articles and rank them in order of their novelty, and their relevance to the participant's practice. The results will be analysed to see whether participants' perception of the novelty of the

articles validate the AI predictions. All data will be collected and stored anonymously.

Informed consent

The study assumes that if you have voluntarily clicked on the link, read this information about the study, and completed the questionnaire, you have consented to participate in the research. The anonymous data collection/storage mean that it will not be possible to remove your data from the study after you have pressed the "Submit" button. No data will be collected if you do not press the "Submit" button.

Confidentiality

No personal data will be collected during the study. The only information collected will be participants ranking of the novelty and relevance of the articles. The information will be collated, analysed, and submitted for publication and/or presentation as appropriate.

Storage of the data

The data will be securely stored on servers at James Cook University for 7 years.

Risks and benefits of the research

The risks of participating in the research appear low. The potential benefits of the research include the development of an empirical basis for the selection of content in CPD and curriculum development that may help identify and reduce biases which have influenced medical education and professional practice in the past.

Results of the study

While the anonymous nature of the data collection prevents distribution of the results to participants, you can contact the principal investigator at andrew.amos@jcu.edu.au for inclusion on a distribution list to receive a copy of the manuscript produced for submission to journals.

Dissemination

In addition to its inclusion in Dr Amos's PhD, it is intended that the results of the research be written up as a research article, submitted for publication, and presented in a public forum. No personal data will be included - all data will be analysed and presented in the aggregate.

Contacts:

Principal Investigator: Dr Andrew Amos - andrew.amos@jcu.edu.au

Primary Ph.D Supervisor: Professor Bunmi Malau-Aduli - bunmi.malauaduli@newcastle.edu.au

References

1. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli B. (2023a) Identifying emerging topics in the peer-reviewed literature to facilitate curriculum renewal and development. *Current Psychology*.

2. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli B. (2023b) Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps. *MEDINFO2023: The Future is Accessible (Proceedings)*.

If you have any concerns about the ethical conduct of the study, please contact:

Human Ethics, Research office

James Cook University, Townsville, Qld 4811

Phone: (07) 4781 5011 (ethics@jcu.edu.au)

Who should I contact if I have any questions?

If you have any questions about the study, please contact the principal investigator **(Andrew Amos)** and/or the primary supervisor (**Prof Bunmi Malau-Aduli**), whose contact details are provided below.

Principal Investigator: Dr Andrew Amos - andrew.amos@jcu.edu.au
Primary Supervisor: Prof Bunmi Malau-Aduli - bunmi.malauaduli@newcastle.edu.au

Electronic Consent

Please select your choice below. Clicking on the "agree button" indicates that

- You have read the above information
- You voluntarily agree to participate
- You consent to aggregated, anonymous data from the research be included in a PhD, a research article to be submitted for publication, and for presentation in a public forum

◯ Yes, I have read the Participant Information Sheet and consent to participate in this research

◯ No, I do not consent to participate in this research (no questionnaire will be offered)

Questsionnaire Page 2: Presentation of articles for ranking.

Articles for ranking

Please read the following 10 article titles. If the meaning of the title is unclear, abstracts are available by hovering above each title (Results/Methods have been removed in the interest of brevity, except where this obscured meaning). After you have finished reading, you will be asked to rank the articles twice, once in order of their novelty to you (with 1 being most novel and 10 being least novel); and once in order of their relevance to your current practice (with 1 being most relevant and 10 being least relevant):

Title 1:
Features of personality and professional burnout syndrome of physicians - analysis based on questionnaires studies.

Title 2:
Using Natural Language Processing and Machine Learning to Identify Hospitalized Patients with Opioid Use Disorder.

Title 3:
Reactions on 'The effectiveness of monoamine oxidase inhibitors in treatment-resistant depressive disorders in clinical practice; a retrospective open-label study'

Title 4:
Restricted Effect of Cerebral Microbleeds on Regional Magnetic Susceptibility. The impact of cerebral microbleeds.

Title 5:

Patient preferences when searching for clinical trials and adherence of study records to ClinicalTrials.gov guidance in key registry data fields.

Title 6:

Identifying Clinical Risk Factors for Opioid Use Disorder using a Distributed Algorithm to Combine Real-World Data from a Large Clinical Data Research Network.

Title 7:

Mental Health Comorbidity Analysis in Pediatric Patients with Autism Spectrum Disorder Using Rhode Island Medical Claims Data.

Title 8

No need to touch this: Bimanual haptic slant adaptation does not require touch. Three experiments using a slant adaptation paradigm.

Title 9:

Relationship of Neurofilament Light (NfL) and Cognitive Performance in a Sample of Mexican Americans with Normal Cognition, Mild Cognitive Impairment and Dementia.

Title 10

Motor functions: Motor development includes the evolution from reflexive to voluntary and goal-directed motor actions.

→

Questionnaire Page 3: Ranking articles for novelty – page prior to ranking.

Articles for ranking

Please rank the following 10 article titles in order of their novelty to you by dragging them from the left box and dropping them in order into the box on the right.

Hovering above the title will show the abstract.

Items	Rank articles in order of novelty (1/top most novel; 10/bottom least novel)
Features of personality and professional burnout syndrome of physicians - analysis based on questionnaires studies.	
Using Natural Language Processing and Machine Learning to Identify Hospitalized Patients with Opioid Use Disorder.	
Reactions on 'The effectiveness of monoamine oxidase inhibitors in treatment-resistant depressive disorders in clinical practice; a retrospective open-label study'	
Restricted Effect of Cerebral Microbleeds on Regional Magnetic Susceptibility.	
Patient preferences when searching for clinical trials and adherence of study records to ClinicalTrials.gov guidance in key registry data fields.	
Identifying Clinical Risk Factors for Opioid Use Disorder using a Distributed Algorithm to Combine Real-World Data from a Large Clinical Data Research Network.	
Mental Health Comorbidity Analysis in Pediatric Patients with	

Autism Spectrum Disorder Using Rhode Island Medical Claims Data.

No need to touch this: Bimanual haptic slant adaptation does not require touch.

Relationship of Neurofilament Light (NfL) and Cognitive Performance in a Sample of Mexican Americans with Normal Cognition, Mild Cognitive Impairment and Dementia.

Motor functions: Motor development includes the evolution from reflexive to voluntary and goal-directed motor actions.

→

Questionnaire Page 4: Ranking articles for novelty – page after ranking.

Articles for ranking

Please rank the following 10 article titles in order of their novelty to you by dragging them from the left box and dropping them in order into the box on the right.

Hovering above the title will show the abstract.

Items	Rank articles in order of novelty (1/top most novel; 10/bottom least novel)
	1 Features of personality and professional burnout syndrome of physicians - analysis based on questionnaires studies.
	2 Mental Health Comorbidity Analysis in Pediatric Patients with Autism Spectrum Disorder Using Rhode Island Medical Claims Data.
	3 Motor functions: Motor development includes the evolution from reflexive to voluntary and goal-directed motor actions.
	4 No need to touch this: Bimanual haptic slant adaptation does not require touch.
	5 Relationship of Neurofilament Light (NfL) and Cognitive Performance in a Sample of Mexican Americans with Normal Cognition, Mild Cognitive Impairment and Dementia.
	6 Restricted Effect of Cerebral Microbleeds on Regional Magnetic Susceptibility.
	7 Identifying Clinical Risk Factors for Opioid Use Disorder using a Distributed Algorithm to Combine Real-World Data from a Large Clinical Data Research Network.

8 9	Patient preferences when searching for clinical trials and adherence of study records to ClinicalTrials.gov guidance in key registry data fields. Reactions on 'The effectiveness of monoamine oxidase inhibitors in treatment-resistant depressive disorders in clinical practice; a retrospective open-label study'
10	Using Natural Language Processing and Machine Learning to Identify Hospitalized Patients with Opioid Use Disorder.

→

Questionnaire Page 5: Ranking articles for relevance. Page after ranking.

Please rank the following 10 articles in order of their relevance to your practice by dragging them from the left box and dropping them into the box on the right, with the most relevant articles at the top, and the least relevant articles at the bottom. Hovering above the title will show the abstract:

Items

Rank articles in order of clinical relevance (1/top most relevant; 10/bottom least relevant)

1. Using Natural Language Processing and Machine Learning to Identify Hospitalized Patients with Opioid Use Disorder.
2. Features of personality and professional burnout syndrome of physicians - analysis based on questionnaires studies.
3. Restricted Effect of Cerebral Microbleeds on Regional Magnetic Susceptibility. The impact of cerebral microbleeds.
4. No need to touch this: Bimanual haptic slant adaptation does not require touch.
5. Motor functions: Motor development includes the evolution from reflexive to voluntary and goal-directed motor actions.
6. Relationship of Neurofilament Light (NfL) and Cognitive Performance in a Sample of Mexican Americans with Normal Cognition, Mild Cognitive Impairment and Dementia.
7. Identifying Clinical Risk Factors for Opioid Use Disorder using a Distributed Algorithm to Combine Real-World Data from a Large Clinical Data Research Network.
8. Mental Health Comorbidity Analysis in Pediatric Patients with Autism Spectrum Disorder Using Rhode Island Medical Claims Data.

9 Patient preferences when searching for clinical trials and adherence of study records to ClinicalTrials.gov guidance in key registry data fields.

10 Reactions on 'The effectiveness of monoamine oxidase inhibitors in treatment-resistant depressive disorders in clinical practice; a retrospective open-label study'

→

Chapter 9: Discussion/implications

Modern medical practice is built upon the accumulated knowledge of clinicians and researchers with roots back to the 19th century. Each individual doctor understands a small fraction of this accumulated knowledge and applies it to an idiosyncratic mix of clinical work with a unique group of patients combined with other medical roles including administration, research, and education.[1,2]

Medical curricula, and related instruments such as frameworks for the maintenance of professional competence like continuing professional development (CPD), are typically created by small groups of experts within specialist areas of practice.[3] Until recently, this reliance upon experts was necessary because of the sheer volume of information and the lack of empirical tools for summarising and analysing the complete set of medical information.[4] No individual doctor could be expert across all domains, and there were no analytic tools capable of providing empirical information about medical knowledge at that level of abstraction.

While the reliance on experts for the construction of medical curricula in developed countries has been successful, at least as measured by ongoing improvements in health indicators such as lifespan,[5] it can be associated with problems, such as expert bias.[6] A commonly cited example is the mistaken assumption by primarily male senior physicians and health researchers that the exclusion of female subjects from heart disease research would not affect treatment outcomes for women.[7,8]

This thesis was designed to develop and validate machine learning (ML) techniques capable of providing empirical support for the processes of medical curriculum development. As stated in section 1.8.1, it:

1) began with an evaluation of the evidence that biases exist that prevent under-represented minorities, including women, from entering specialist medical training;

2) trained a self-organising map (SOM) to represent the complete domain of medical knowledge contained within high-quality peer-reviewed medical journals indexed by the Medline database (MedSOM);
3) applied the MedSOM to reveal the structure of knowledge contained in a core psychiatric textbook;
4) used a bibliometric algorithm to identify emerging topics in the medical literature suitable for consideration for inclusion in medical curricula;
5) validated the emerging topics identified by a bibliometric algorithm for their suitability for guiding the continuing professional development of practicing psychiatrists; and
6) described technical improvements in the SOM algorithm required to apply the SOM technique to the entire Medline database instead of the small subset of articles between 2004-2008.

To address these aims, the thesis hypothesised that:

1) there are systematic biases in the methods of selecting junior doctors into specialist training programmes associated with the under-representation of certain groups including women and ethnic minorities;
2) the structure and relationships of the medical knowledge contained within the complete set of articles indexed by the Medline database using a controlled vocabulary of Medical Subject Headings (MeSH) can be summarised in human-intelligible form by a SOM;
3) the knowledge represented by the reference lists of a core psychiatric textbook can be interpreted by a SOM trained on the Medline database;
4) topics of emerging interest in the peer-reviewed medical literature indexed by the Medline database can be identified using a bibliometric algorithm;

5) topics of emerging interest in the peer-reviewed medical literature identified by a bibliometric algorithm will be viewed as novel and relevant to practice by consultant psychiatrists; and
6) improvements in the hardware and SOM algorithms will make it feasible to train the entire corpus of the Medline database instead of a subset.

Table 9-1 outlines the primary findings from chapters 3 to 8 and describes how these findings relate to the aims and hypotheses of the thesis.

Table 9-1: Primary findings of chapters 3 to 8 and their contribution to the thesis aims and hypotheses

Chapter	Major Findings	Contributions to the thesis
3[8]	The systematic review found 35 articles which used empirical measures to detect evidence of bias preventing the selection of under-represented minorities into medical specialist training. In order of prevalence, they addressed the under-representation of women (21/35), international medical graduates (10/35), and race/ethnicity (9/35). Other than a small group of well-powered studies of selection into general practice training in the UK the reviewed studies were methodologically questionable, providing only preliminary evidence that bias in selection measures might contribute to the under-representation of some groups.	There is no doubt that systematic biases exist in the practice of medicine, but the existing evidence is not designed or powered to effectively detect, and therefore cannot reliably exclude the possibility of systematic biases that prevent the selection of under-represented minorities into medical specialist training. Given the ongoing evidence of gender gaps in areas like the seniority of authors across articles published in prestigious medical journals, an independent empirically derived model of the medical knowledge contained in the peer reviewed medical literature has promise as a tool for the identification and reduction of such biases.
4[9]	This article reported a refined version of the batch algorithm which successfully trained a medical SOM (MedSOM) on the complete set of more than 30 million Medline articles and the complete set of medical subject headings (29,917 MeSH) available at 01/01/2021, compared	This article provided an existence proof that a SOM can represent a human-intelligible visualization of the entire corpus of articles indexed by the Medline database, and reveal meaningful structural features, including the differentiation of and relationship between medical and

	with the 2 million articles and 2,300 MeSH used to train the Skupin SOM. Medical and psychiatric MeSH were differentiated by colour, revealing both the internal relationships within each domain, as well as their interaction with the other domain.	psychiatric knowledge. In terms of the ultimate goal of producing a map of medical knowledge capable of supporting medical curriculum development, this can be viewed as a necessary baseline establishing general features and organisation, like a simple list of placenames on the face of a political map.
5[10]	This article reported the results of applying a bibliometric algorithm to the psychiatric literature indexed by the Medline database to identify topics of emerging importance. Emerging topics were identified across nine sequential five-year periods (demi-decades) between 1972-1976 and 2012-2016 and visualized using co-word analyses. Themes arising from the visualisations of the most common emerging topics in each period were consistent with known landmarks in psychiatric/psychological history, such as the sequential release of new editions of the Diagnostic and Statistical Manual of Mental Disorders.	This article demonstrated that it is feasible to identify emerging topics based on the characteristics of articles themselves, rather than the pattern of citations they generate. This has the advantage of identifying emerging topics, rather than important articles, in addition to avoiding the delay associated with waiting for citation patterns to emerge, which can take several years to become apparent.

6[11]	Much of the power of machine learning algorithms is the ability to detect patterns without specifying in advance what type of patterns are expected. The article appended as chapter six reported technical innovations in the batch self-organising map algorithm that allowed a medical SOM (MedSOM) to include the entire set of more than 30 million articles, and the entire set of 29,917 MeSH, instead of the small subset reported in previous work such as Skupin et al.[12]	In order to develop maps of medical knowledge capable of supporting medical curriculum development it is necessary to be able to consider the entire corpus of articles contained within medical databases like Medline. Substituting a batch SOM algorithm for a dense SOM algorithm led to processing times hundreds of times faster, and using a small fraction of the memory. Technical improvements to the batch SOM algorithm led to more incremental improvements of 5-10% of processing speed and reducing the memory requirements by 50%. The combination of both allowed for the application of the SOM to the entire corpus of Medline articles. Ongoing technical improvements will be necessary to extend the utility of the SOM to functions such as continuous temporal modelling of medical and psychiatric knowledge spaces.
7[13]	The medical SOM (MedSOM) trained in previous chapters was used to interpret reference lists from ten editions of a core psychiatric textbook. MedSOM coherently clustered references into six	The MedSOM coherently organized the reference lists of a core psychiatric textbook and differentiated them into meaningful clusters based on psychiatric subspecialties. This demonstrates that the

	psychiatric knowledge domains across ten editions between 1967 and 2017. Clustering was evident at the level of broad psychiatric specialties including Adult Psychiatry, Child Psychiatry, and Administrative Psychiatry.	organization of the 2-dimensional knowledge space represented by the MedSOM maps coherently on to the knowledge space represented by that psychiatric textbook, supporting the validity of its use for curriculum development.
8[14]	To test the validity of the emerging topics identified in chapter 5 a sample of practicing psychiatrists were asked to rank the novelty and relevance to practice of 5 Medline indexed articles with at least one emerging topic, and 5 Medline indexed articles with no emerging topics. The psychiatrists ranked the emerging-topic articles as significantly more novel and more relevant than non-emerging topic articles. Novelty and relevance of individual articles were not highly correlated.	This article demonstrated that a bibliometric algorithm can identify articles that practicing psychiatrists rank as both novel and relevant to their practice. This supports the potential validity of using bibliometric algorithms to identify materials that should be considered for inclusion in medical curricula based on their novelty and practical relevance.

9.1 - Developing a hybrid epistemology of machine learning algorithms

As discussed in Chapter 2, the main barrier to the use of ML algorithms for complex medical tasks like curriculum development is the difficulty of understanding what their outputs mean. Even where those outputs can be directly verified against gold standard assessments, such as the biopsy of imaging-identified lesions, doctors can be reluctant to rely upon algorithms they do not understand.[15]

Babushkina & Votsis point out that if you consider the decision-making process of a human expert relying in part upon the product of an artificial intelligence, you can analyse the reasoning chain into machine and human claims and inferences. They give the example of a human expert integrating into their diagnostic process and treatment decisions the advice of an AI algorithm which, based on an x-ray, estimates a patient's probability of having pneumonia as 85%, of having cancer as 10%, and of having tuberculosis as 5%. However, Babushkina & Votsis note the pattern recognition performed by AI algorithms that analyse images is driven by the statistical similarity between a novel image and the features of a set of training images rather than the distribution of images of that kind between patients with and without a particular diagnosis. They describe this type of analysis as trivial similarity to contrast it with the meaningful similarity human experts perceive between cases.[16]

As a result, it is a mistake for the human expert to interpret the probabilities generated by the AI as equivalent in meaning to the probabilities estimated by a human expert interpreting the same x-ray. Instead, the human expert must be aware of and act upon the knowledge that the AI algorithm in this case has no understanding of anything other than the pure visual representation in the presented x-ray, and that it is the expert's responsibility to consider that evidence in the context of all the other information they have about the patient's presentation such as symptoms, history, treatment, and physical examination.[16]

In contrast to the example of an AI algorithm estimating the probability of different diagnoses based on imaging studies, it is not possible to provide a gold standard to define the meaning of ML

algorithms like MedSOM which visualize features of the knowledge contained within incomprehensibly large sets of data about medical research. As a result, it is necessary to validate their meaning using other techniques.

The most important qualities for a map of medical knowledge intended to be useful for developing medical curricula include the capacity to summarise the entire body of published medical research in terms of meaningful units such as the phrases of a controlled vocabulary; to accurately differentiate between distinct elements of the information contained within that research; to represent the relationships between those distinct elements; and to have the capacity to make comparisons between them based on relevant characteristics such as their novelty and clinical relevance.

The articles comprising this thesis establish a baseline map of medical knowledge and begin the task of validating its meaning against these desirable qualities. Most completely, Chapter 4 established that MedSOM has the capacity to summarise the entire body of medical research published in the publicly available Medline database, and to differentiate medical and psychiatric information structures.

Chapter 7 demonstrated that MedSOM can organise the reference lists of a psychiatric textbook into clusters consistent with standard psychiatric specialties such as Adult Psychiatry, Child Psychiatry, and Administrative Psychiatry. Supplementary material 2 to Chapter 7 discussed how subsets of the Medline database can be superimposed on the MedSOM to identify concepts and articles relevant to some topic of particular interest. Supplementary material 2 used the research of Song and colleagues to illustrate how information about sex and gender specific health could be integrated into a medical curriculum using MedSOM.[17]

Chapter 5 described how a bibliometric algorithm could be used to identify topics of emerging interest based on sudden increases in the number of articles associated with medical subject headings. In the same way as information about sex and gender specific health could be integrated into a medical curriculum, so the output of the emerging topics algorithm could be superimposed

upon MedSOM to identify areas of knowledge and articles containing novel information likely to be relevant to practice.

Finally, Chapter 8 established that articles selected by the bibliometric algorithm described in Chapter 5 were ranked as more novel and more relevant to practice by psychiatrists required to engage in continuing professional development. This demonstrated that the type of hybrid epistemology comprising a chain of inferences integrating AI products and human decision-making is entirely plausible using the MedSOM and the emerging topics algorithm.

The explainable artificial intelligence (xAI) field has started to establish frameworks for optimally integrating the outputs of AI algorithms and human experts using the sort of hybrid epistemologies described by Babushkina & Votsis.[16] Kumar and Sharma argue that integration of algorithm and human expert achieves better performance than either alone.[18]

Perhaps the most systematic work on safely integrating AI products with human experts has been done by Sanneman and Shah, who emphasise the importance of clearly defining the intended purpose, the role of AI in achieving that purpose, and the features of the human-AI interaction likely to affect performance, including the cognitive workload of the task, the complexity of the AI product, and the human's level of trust in the AI product.[19]

9.2 - Applying a hybrid epistemology of machine learning algorithms to medical curriculum development

While the research reported in this thesis takes the first steps towards a map of medical knowledge capable of supporting medical curriculum development, further steps are needed. This research has established that a map of the entire Medline corpus is feasible; that the map represents a knowledge structure that is consistent with a core psychiatric textbook; and that it has the potential to be used to provide a baseline for understanding the outputs of other ML products such as the emerging topics algorithm, in the same way that political maps showing city locations provide context for understanding the relative location of geographic features like mountains.

However, to be useful as an input into medical curriculum development, it would need to be extended in specific ways. Among other tasks, curriculum development requires the specification of a set of knowledge, skills, and attributes (KSAs) required for competent medical practice. These are often subdivided into stages of training, with more basic KSAs acquired early and more advanced KSAs acquired later; and into specialist areas of practice. For example, physicians and surgeons share a set of generalist KSAs such as basic anatomical knowledge, the principles of medical therapeutics, and standard methods for assessment.

In addition, the maintenance of medical curricula requires that obsolete KSAs be removed, while important new KSAs are added. The evolution of the set of KSAs underlying competent medical practice can be considered both as individual decisions to add or remove items at points in time, or as the cumulative result of those decisions over time.

While Chapter 4 established that MedSOM can differentiate between psychiatric and medical concepts, the differentiation was based on the controlled vocabulary created by the National Library of Medicine (NLM)[20] rather than the KSAs of physicians and psychiatrists respectively.

Therefore, to make MedSOM useful for curriculum development it would be necessary to develop techniques to differentiate the KSAs associated with the various medical specialties, rather than relying upon the NLM's controlled vocabulary. An obvious way to achieve this would be to train or modify MedSOM by specialty using the subset of articles published in specialty-specific journals.

It may prove more difficult to differentiate the KSAs appropriate for different stages of training (and, by extension, different stages of practice such as medical school, internship, registrar training, and consultancy). One possible approach would be to filter the article sets used to train MedSOM using information obtained from other datasets. For example, the electronic medical record systems used at some hospitals record detailed information about the types of medical activities most commonly performed by doctors at different stages of training. Other possibilities include the text of existing

medical curricula, and the text of questions used for assessment of doctors at different stages and training for different specialties.

Conceptually, the extension of MedSOM to accommodate changes over the stages of training is a specific example of the general capacity for MedSOM to respond dynamically to variations in the underlying data set. The most natural example would be the facility to accommodate changes in the patterns of articles published over time. While Chapter 7 examined the evolution of the patterns of references over subsequent editions of a core psychiatric textbook, which de facto involves the interpretation of changes over time, the underlying MedSOM was trained on the complete set of articles, not period-specific articles.

Perhaps the most potentially useful hybrid epistemology arising from the research described in this thesis would involve the combination of the emerging topic algorithm and MedSOM. The emerging topic algorithm could identify the most important new topics; MedSOM could link those topics to areas of the medical knowledge domain, refined by characteristics such as area of specialty; and the highest impact recent articles in those areas could be presented to a human expert to consider for inclusion in the curriculum.

9.3 - The complementary strengths of self-organising maps and large language models

Over the course of the research comprising this PhD the fields of AI and ML have been transformed by the emergence of powerful new models grouped under the title of large language models (LLMs). LLMs leveraged long-established neural network techniques including reinforcement learning, backward-propagation, and forward-propagation, with the rapidly advancing computational power of modern GPU arrayed in huge data centres, to provide previously unimaginable AI services.[21]

The power of LLMs is largely due to a core class of algorithms called transformers, which can learn the syntactic and semantic properties of enormous datasets and respond to questions expressed in everyday language with highly detailed and meaningful answers. Although the exact mechanisms are not precisely known, the most common type of LLM (text-to-text), which responds to questions

expressed in written text with answers in written text can be thought of as a statistical function where the algorithm probabilistically selects each new word (or, more precisely, each new unit of meaning) of an answer based on the set of previous words presented to and output by the algorithm.[22]

Despite the enormous power of LLMs like ChatGPT, they are more complementary to than competitive with SOMs for the purpose of visualising the structures and relationships of the domains of medical knowledge. There are very few published examples of maps of medical knowledge generated by LLM algorithms approaching the coverage of the MedSOM, and their role in map generation is completely different.

The most prominent example of an LLM-generated map of medical knowledge is one of the first, reported by González-Márquez et al. in 2024.[23] They used the transformer-based LLM algorithm PubMedBERT to generate a map (the GM map hereafter) based on the ~21 million articles in the Medline database with abstracts written in English. The GM map can be distinguished from MedSOM both in the type of data transformation being performed, and by the outputs provided.[23]

MedSOM relied upon the controlled vocabulary of MeSH curated by the NLM to define its units of meaning, representing each article in the Medline database as a vector of ~10-20 MeSH.[9] Rather than adopt the fixed and readily understood MeSH, the GM map trained the PubMedBERT LLM algorithm on the 21 million English abstracts and extracted a vector of 768 elements from a hidden layer of the algorithm to represent each article. It then applied the dimension reduction algorithm called t-distributed stochastic neighbour embedding (t-SNE) to the set of LLM derived vectors to produce a two-dimensional map of all the Medline articles with English abstracts (Figure 9-1).[23]

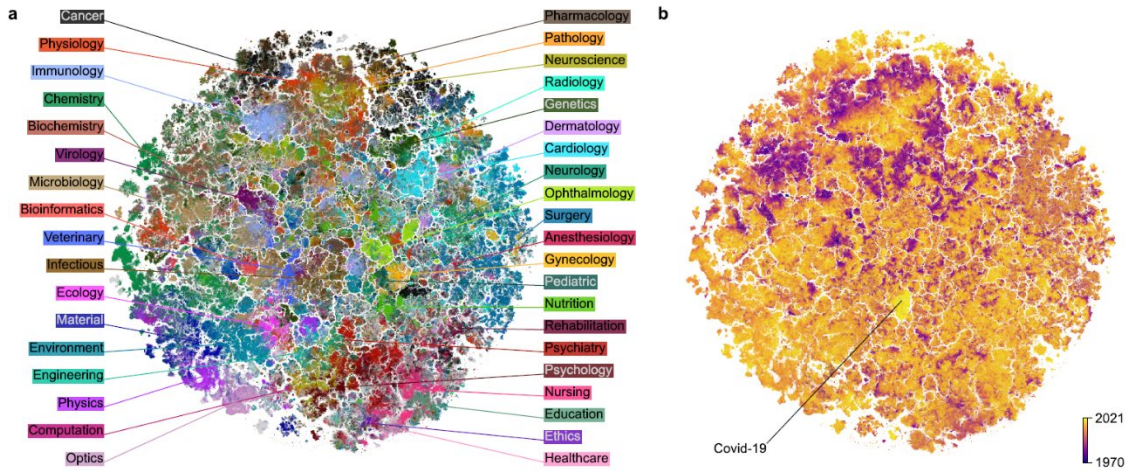


Figure 9-1: González-Márquez et al.'s map of the medical literature[23] is licensed under CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/)

Figure 9-1 shows that the output of the GM map is quite different from the MedSOM output (Figure 4-1, page 105). Each point on the GM map represents a single article, while each point on the MedSOM map represents a node with an abstract meaning defined by a vector indicating a particular mix of the 29,712 MeSH in the NLM's controlled vocabulary.[23] Less technically, the GM map represents the position of individual articles and their relationship to other articles, while the MedSOM nodes define a region of knowledge in terms of abstract categories with defined meanings (MeSH).

The only information about abstract categories of meaning shown by the GM map are 38 categories of medical knowledge (shown in two columns of highlighted text either side of the map in Fig 9-1(a)). These abstract categories were attributed to each article if the category appeared in the title of the article. For example, an article was categorised as being in the "Cancer" category if the word "cancer" appeared in the title.[23]

This comparison illustrates some of the known strengths and limitations of SOMs compared with other methods of dimension reduction (like t-SNEs). SOMs have a fixed architecture defined by a certain number of nodes and a deterministic algorithm that defines regions of a two-dimensional

map in terms of abstract meanings grounded in a controlled vocabulary. They have relative strength in mapping the structure and relationships of medical knowledge at a high level of abstraction (the global structure of the map). The creation of a map using the combination of an LLM with t-SNE have a fluid architecture created by a combination of probabilistic and deterministic algorithms which have relative strength in mapping the structure and relationships of medical knowledge at the specific level of individual articles (the local structure of regions of the map).

There is an obvious complementarity between the processes and outputs of the MedSOM and the GM map, suggesting that it might be possible to combine these approaches to deliver the best features of both. Which approach or combination of approaches is an open question which remains to be empirically explored. It would be particularly interesting to leverage the MedSOM's shape and fixed meanings with the t-SNE's capacity to organise the relationship between individual articles, and LLM's capacity to translate abstract meaning into human intelligible text. The substitution of the 29,712 element MeSH vectors of the MedSOM and the 768 element vectors taken from a hidden layer of an LLM for the GM map would be a feasible starting point for exploring the integration of these disparate algorithms.

9.4 - Strengths and limitations of the research

In common with the field of ML algorithms itself, the strengths and limitations of the current research are intimately linked. The great strength of ML algorithms like SOMs is their ability to analyse enormous datasets of any level of complexity and condense high dimensional patterns into 2-dimensional maps that describe patterns that are intelligible to human beings. SOMs also do not require pre-specification of the patterns of interest – rather they detect and expose the patterns that exist, including patterns that are not tangible to human beings in their high dimensional forms.

SOMs have several limitations directly related to these strengths. While they can detect previously unknown and intangible patterns, the mechanisms by which they detect those patterns are not precisely known, and the meaning of the patterns themselves usually requires interpretation or

translation. A significant part of the thesis was involved in searching for novel ways to validate the meaning of MedSOM by demonstrating that its structure coherently organised related materials, and that the products of the emerging topics algorithm were meaningful to a group of practicing psychiatrists.

Another strength and complementary limitation of the research was its exploratory nature. The application of SOMs to the complete set of a medical research database as the first step towards a tool for use in medical curriculum development is novel, but for that reason it is difficult to evaluate against similar research. Given the potential value of an empirically derived map of medical knowledge that summarises information about the entire set of peer-reviewed medical research indexed by Medline, and the absence of any other well-established techniques for providing a comprehensive empirical basis for curriculum development decisions, this appears a reasonable trade-off.

Finally, the thesis provides limited evidence about the integration of the products of ML algorithms into human experts' decision-making processes. The single experiment involving human subjects involved a small set of subjects performing a simple ranking task with anonymous recruitment through a professional body utilising an online questionnaire. This is a reasonable pilot study, and it would be possible to significantly improve the design and increase the effect size by targeting doctors from a specific medical or psychiatric sub-specialty and presenting them with treatment and control articles from that sub-specialty.

9.5 - Reflections

Implementation of this course of research has required the acquisition of a broad range of technical, statistical, and experimental skills. While the author was familiar with traditional object oriented and functional programming languages, the computational demands of training SOMs on datasets with millions of items characterised by tens of thousands of dimensions necessitated the use of parallel programming algorithms instantiated in the technically complicated C++ language.

While it generated a relatively small part of the thesis, overcoming these technical challenges consumed the greatest amount of time over the course of the project. Rather serendipitously, the programming skills learned during this process became far more professionally salient after the explosion of interest in artificial intelligence which occurred part-way through the thesis.

Another unexpected outcome of the research program was the need to constantly recalibrate expectations. While training a SOM appeared conceptually simple at the outset, interpreting the results emphasised the lessons of the explainable artificial intelligence (xAI) field, and demanded the creation of several novel extensions to validate the meaning of the results.

Perhaps most importantly, investigation of the xAI field led to the concept of a hybrid epistemology comprising both human and ML components. Above and beyond the xAI frameworks for interpreting and explaining the mechanisms and outputs of ML processes, the insight that it is a mistake to think that ML products have the same meaning as human assertions with similar structures was crucial.

9.6 - References

1. Caddick ZA, Fraundorf SH, Rottman BM, Nokes-Malach TJ. Cognitive perspectives on maintaining physicians' medical expertise: II. Acquiring, maintaining, and updating cognitive skills. Vol. 8, Cognitive Research: Principles and Implications. Springer Science and Business Media Deutschland GmbH; 2023.

2. Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc [Internet]. 2011;122(319):48–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21686208%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3116346

3. Harden RM. AMEE Guide No. 21: Curriculum mapping: A tool for transparent and authentic teaching and learning. Med Teach. 2001;23(2):123–37.

4. Larsen PO, von Ins M. The rate of growth in scientific publication and the decline in coverage provided by science citation index. Scientometrics. 2010;84(3):575–603.

5. Australian Bureau of Statistics. Life expectancy, 2021-2023 [Internet]. ABS Website. 2025 [cited 2025 Mar 5]. Available from: https://www.abs.gov.au/statistics/people/population/life-expectancy/latest-release

6. Ibrahim H, Juve AM, Amin A, Railey K, Andolsek KM. Expanding the Study of Bias in Medical Education Assessment. J Grad Med Educ. 2023 Dec 1;15(6):623–6.

7. The Lancet. Cardiology's problem women. The Lancet [Internet]. 2019;393(10175):959. Available from: http://dx.doi.org/10.1016/S0140-6736(19)30510-0

8. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Systematic review of specialist selection methods with implications for diversity in the medical workforce. BMC Med Educ. 2021 Dec 1;21(1).

9. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli B. Defining the boundaries of psychiatric and medical knowledge: applying cartographic principles to self-organising maps. Stud Health Technol Inform [Internet]. 2024 [cited 2025 Jul 15];310:795–9. Available from: https://pubmed.ncbi.nlm.nih.gov/38269918/

10. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Identifying emerging topics in the peer-reviewed literature to facilitate curriculum renewal and development. Current Psychology [Internet]. 2022 Dec 12;42:30813–24. Available from: https://link.springer.com/10.1007/s12144-022-04090-y

11. Amos A, Lee K, Sen T, Bunmi G, Malau-Aduli S. Novel sparse matrix algorithm expands the feasible size of a self-organizing map of the knowledge indexed by a database of peer-reviewed medical literature. Neural Comput Appl. 2025;Submitted.

12. Skupin A, Biberstine JR, Börner K. Visualizing the Topical Structure of the Medical Sciences: A Self-Organizing Map Approach. PLoS One. 2013;8(3).

13. Amos AJ, Lee K, Sen Gupta T, Malau-Aduli BS. Validating the knowledge represented by a self-organizing map with an expert-derived knowledge structure. BMC Med Educ. 2024;24(416).

14. Amos AJ, Lee J, Sen Gupta T, Malau-Aduli BS. Testing the validity of emerging topics in psychiatry identified by a machine learning algorithm with a sample of working psychiatrists. Comput Biol Med. 2025;Submitted.

15. Mohseni S, Zarei N, Ragan ED. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans Interact Intell Syst. 2021 Aug 31;11(3–4):1–45.

16. Babushkina D, Votsis A. Epistemo-ethical constraints on AI-human decision making for diagnostic purposes. Ethics Inf Technol. 2022 Jun 1;24(2).

17. Song MM, Simonsen CK, Wilson JD, Jenkins MR. Development of a PubMed based search tool for identifying sex and gender specific health literature. J Womens Health. 2016 Feb 1;25(2):181–7.

18. Kumar P, Sharma M. Data, Machine Learning, and Human Domain Experts: None Is Better than Their Collaboration. Int J Hum Comput Interact. 2022;38(14):1307–20.

19. Sanneman L, Shah JA. The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. Int J Hum Comput Interact. 2022;38(18–20):1772–88.

20. National Library of Medicine. NLM - Introduction to MeSH [Internet]. 2019 [cited 2019 Aug 27]. Available from: https://www.nlm.nih.gov/mesh/introduction.html

Chapter 10: Conclusions and Future Research

Based on the research conducted in the course of this thesis, it is concluded that it is possible to create an empirically derived map that summarises the content, organisation, and structure of all the knowledge indexed by the Medline database of peer-reviewed medical literature (MedSOM). It has also been possible to provide preliminary support for the validity of this map by using its structures to differentiate between primarily medical and primarily psychiatric concepts as they appear in the literature, and to analyse the knowledge structures contained within a core psychiatric textbook.

In addition, it can be concluded that bibliometric algorithms can be developed which extract useful information about the medical subject headings which comprise the knowledge units mapped by the MedSOM. It was demonstrated that this information meaningfully differentiated novel and relevant articles for a group of practicing psychiatrists.

This research starts the first steps towards an empirically derived map of medical knowledge capable of informing academic medical activity such as medical curriculum development. An example of the potential uses of such a map is the reduction of systematic biases evident in features like the gender gap in senior authors on articles in significant medical research journals.[1]

10.1 - Expanding on the meaning and validity of MedSOM

The exploratory nature of the current research means that there are many potential extensions of MedSOM and the emerging topics algorithm that could be pursued to facilitate medical curriculum development. If MedSOM is considered as a basic map of a knowledge domain, like a political map characterised only by the placenames of important landmarks like cities and countries, then the most important extension would be to develop the capacity to represent changes over time.

In terms of a political map, this could be achieved as a series of snapshots at points in time showing all the cities and countries and their locations at the beginning of each period of time. A more dynamic approach would be to show the evolution of political entities with an animated display that

showed the appearance/disappearance of cities/countries, for example the founding of the USA in the 18th century or the end of the USSR at the end of the 20th century.

For the purposes of curriculum development, MedSOM could be extended using Denny's ReDSOM technique, which sequentially trains a SOM on different periods of data. This would produce a series of snapshots in time, representing the knowledge published over a particular period.[2] No existing SOMs provide a dynamic representation of the changes in the knowledge structures over time rather than at points in time. However, it may be possible to add such a dynamic function to MedSOM by supplementing it using explainable artificial intelligence techniques which interrogate existing AI algorithms with supplementary algorithms that make inferences about qualities not considered by the original algorithms.[3]

Another potential extension of MedSOM that would be valuable for curriculum development was discussed in Supplementary Material 2 to Chapter 7. Again, if MedSOM is considered as a basic map of the domain of medical knowledge like a map of placenames, then it can be leveraged by superimposing maps representing other forms of knowledge. The supplementary material gave the example of superimposing information about sex and gender specific health (SGSH) on the MedSOM to detect and correct for the history of ignoring differences between the sexes due to practices such as excluding women from medical trials due to the potential for pregnancy.

10.2 - Integrating MedSOM into the process of medical curriculum development

Perhaps the greatest potential for improving the utility of MedSOM and the emerging topics algorithm for medical curriculum development and training purposes would arise from calibrating the map using overlapping sets of categorical and individual information. For example, it would be possible to calibrate MedSOM's topology by superimposing information from other data sets such as medical specialties, and individual interests or history. The example of filtering MedSOM's results by considering only journals from specific medical specialties has already been discussed, but this could be leveraged by adding information about an individual clinician's preferences within that specialty.

With reference to the need for a hybrid epistemology considered in Chapter 9, this type of calibration of MedSOM would be extremely useful for integrating MedSOM's outputs into the process of curriculum development. It would allow human experts to interrogate MedSOM for specific information, for example by focusing on the knowledge structures within specific fields of medicine; comparing the relative importance of those fields in terms of metrics like number of publications, novelty of results, and relevance to clinical practice; and providing an empirically derived standard to guide particular curriculum development activities such as distributing training experiences across all specialties in a way proportional to the importance of those activities at different stages of training and practice.

10.3 - Application of MedSOM to the identification and reduction of bias

As discussed in Supplementary Material 2 to Chapter 7, the capacity to superimpose other sources of information on MedSOM holds great promise for identifying and reducing sources of bias reflected in the published peer-reviewed medical literature. The research by Brück discussed in Chapter 2 found that gender gaps regarding senior authors in prestigious medical journals were partially explained by the less strategic use of keywords by female authors.[1] This illustrates that many biases can be detected by analysing the linguistic features of written scientific products like journal articles.

An empirically derived map of medical knowledge like MedSOM can be useful for identifying and reducing biases in medical curricula, training, and related fields like research by establishing the organisation and structures of the knowledge domains in which biases occur. For example, if Brück's work is taken as a starting point to suggest that there is a gender gap in the rate of senior authorships for women across all medical specialties,[1] MedSOM could be used to analyse whether the nature of the gap is different across different domains of medical knowledge, across different periods of publication, and across other categories such as patient population, diagnostic group, or demographic features.

Acknowledging the power and limitations of ML algorithms like SOMs, and the wide variety of potential extensions of the current research, two priorities should be the extension of the static map to accommodate changes in time, and the superimposition of information about emerging topics onto MedSOM to allow for the identification of emerging topics across topics.

Denny's ReDSOM model is the most promising prototype for a temporally dynamic map of medical knowledge, but other possibilities exist.[2,4–6] Most of the work in this area has considered time series data[7–9] rather than the evolution of a knowledge map over time. There have also been some very ambitious attempts to model the evolution of science over time, but their complexity tends to reduce their value for practical applications like curriculum development.[6,10]

As discussed, a primary consideration for the use of ML algorithms like MedSOM for the identification and reduction of biases in medical training, education, and professional development is to reconcile the meaning of ML outputs with the meanings attributed to them by human experts. It is likely that significant experimental work will need to be done to achieve this goal.

10.4 - References

1. Brück O. A bibliometric analysis of the gender gap in the authorship of leading medical journals. Communications Medicine. 2023 Dec 1;3(1).

2. Denny, Williams GJ, Christen P. Visualizing temporal cluster changes using Relative Density Self-Organizing Maps. Knowl Inf Syst. 2010;25(2):281–302.

3. Antoniadi AM, Du Y, Guendouz Y, Wei L, Mazo C, Becker BA, et al. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. Applied Sciences (Switzerland). 2021 Jun 1;11(11).

4. Sarlin P, Yao Z. Clustering of the self-organizing time map. Neurocomputing. 2013 Dec 9;121:317–27.

5. Sarlin P. Self-organizing time map: An abstraction of temporal multivariate patterns. Neurocomputing. 2013 Jan 1;99:496–508.

6. Chavalarias D, Cointet JP. Phylomemetic Patterns in Science Evolution-The Rise and Fall of Scientific Fields. PLoS One. 2013 Feb 11;8(2).

7. D'Urso P, De Giovanni L. Temporal self-organizing maps for telecommunications market segmentation. Neurocomputing. 2008;71(13–15):2880–92.

8. Krishnan KJ, Mitra K. A modified Kohonen map algorithm for clustering time series data. Expert Syst Appl. 2022 Sep 1;201.

9. Aguayo L, Barreto GA. Novelty Detection in Time Series Using Self-Organizing Neural Networks: A Comprehensive Evaluation. Neural Process Lett. 2018 Apr 1;47(2):717–44.

10. Lobbé Q, Delanoë A, Chavalarias D. Exploring, browsing and interacting with multi-level and multi-scale dynamics of knowledge. Inf Vis. 2022 Jan 1;21(1):17–37.

Appendix A: Approvals for Human Experimentation

Figure A-1: James Cook University - HREC Approval for Research Involving Human Subjects – H9432

Figure A-2: Approval to distribute survey to RANZCP members from Policy, Practice and Research Committee