MDRP2: Journal Monograph Task

Searching for the needle in the haystack: the application of a literature searching tool to retrieve information to allow optimum classification of MMR variants.

Varun Kaushik
Student ID: 697258
ORCID ID: 0000-0001-5649-4773

Supervisors: Finlay A. Macrae & John-Paul Plazzer
Department of Colorectal Medicine and Genetics
Royal Melbourne Hospital

ABSTRACT

Background:

Pathogenic variants in the mismatch repair (MMR) genes are the drivers of Lynch Syndrome; optimal variant interpretation is required for the management of suspected and confirmed cases. The International Society for Hereditary Gastrointestinal Tumours (InSiGHT) provides expert classifications for MMR variants for the US National Human Genome Research Institute's (NHGRI) ClinGen initiative and interprets variants with discordant classification and those of uncertain significance (VUSs). Given the onerous nature of extracting information related to variants, literature searching tools which harness artificial intelligence may aid in retrieving information to allow optimum variant classification.

Methods:

In this study, we described the nature of discordance in a sample of 80 variants from a list of variants requiring updating by InSiGHT for ClinGen by comparing their existing InSiGHT classifications with the various submissions for each variant on the US National Centre for Biotechnology Information's (NCBI) ClinVar database. To identify the potential value of a literature searching tool in extracting information related to classification, all variants were searched for using a traditional method (Google Scholar) and literature searching tool (Mastermind Genomenon) independently. Descriptive statistics were used to compare: the number of articles before and after screening for relevance and the number of relevant articles unique to either method. Relevance was defined as containing the variant in question as well as data informing variant interpretation.

Results:

A total of 916 articles were returned by both methods and Mastermind averaged four relevant articles per search compared to Google Scholar's three. Of relevant Mastermind articles, 193/308 (62.7%) were unique to it, compared to 87/202, (43.0%) for Google Scholar. For 24 variants, either or both methods found no information. All 6/80 (20%) variants with pathogenic or likely pathogenic InSiGHT classifications have newer VUS assertions on ClinVar.

Conclusion:

Our study demonstrated that for a sample of variants with varying discordant interpretations, Mastermind was able to return on average, a more relevant and unique literature search. Google Scholar was able to retrieve information that Mastermind did not, which supports a

conclusion that Mastermind could play a complementary role in literature searching for classification. This work will aid InSiGHT in its role of classifying MMR variants.

KEYWORDS

'Lynch syndrome,' 'genetic variation, classification,' 'database management systems,' 'information storage and retrieval' 'literature searching.'

1. Introduction

Lynch Syndrome (LS) is the most common aetiology of hereditary colorectal neoplasia with a prevalence of 3-5% amongst colorectal cancer patients. (1) LS is characterised by pathogenic variations in the DNA mismatch repair (MMR) genes (*MLH1, MSH2, MSH6* and *PMS2),* which are highly penetrant and confer an increased risk of colorectal cancer, amongst other cancers. (2) Today, LS is often diagnosed with DNA sequencing technology which can identify known pathogenic MMR variants. With advances in this technology and its widespread use in the diagnosis and management of LS and suspected cases, our awareness of novel variants in the MMR genes has increased dramatically. To deal with this influx, expert groups who specialise in particular syndromes e.g LS, have been formed to optimise classification for novel variants. The International Society for Hereditary Gastrointestinal Tumours (InSiGHT) provides expert classifications on MMR variants. InSiGHT's role has now also expanded to being recognised as a NHGRI's ClinGen Variant Curation Expert Panel (VCEP) addressing the task of curation of MMR variants in the NCBI's ClinVar database. ClinVar was launched in 2013 in an effort to increase variant data sharing and promote standardised classification of variants.

A key role of the InSiGHT VCEP is to reclassify variants on the ClinVar database whose genotype-phenotype relationship is unclear or not definitive as understood by submitters. These include variants of uncertain significance (VUS); in addition many variants receive discordant pathogenicity assignments when submitted to databases such as ClinVar. Discordance is multifactorial and variant interpretation is often generated from multiple sources, leading to a 'silo effect,' whereby information is considered in isolation by different submitters. The result is a lack of centralised, contemporary information pertaining to a particular variant. (3) VUSs pose a particular clinical problem as although they are not identified as benign with reference to the reference human genome reads, a detrimental influence of the function of the gene is not apparent enough for them to be declared pathogenic on the basis of contemporary data. This leaves families carrying these variants in diagnostic limbo. (4) For discordant variants, misclassification can result in serious clinical mismanagement across and within families, especially in the case where a variant is misclassified as benign and was later reclassified as pathogenic. (5)

To classify a variant, a biocurator may face a seemingly never-ending literature search which may return many irrelevant results. Collecting information that is relevant to variant interpretation is an overwhelming manual task which will only become more onerous with the rate at which literature is now produced. However, there are now variant-oriented search systems that could improve the quality of search results and by extension improve the efficiency of the curation process. (6) These literature searching tools are able to find articles that mention specific variants using artificial intelligence and natural language processing. They have been purported to increase the yield of a literature search compared to traditional search methods. (6) Whilst the literature does describe open source tools such as tmVar2.0 and LitVar, which have been demonstrated to yield more articles than a standard PubMed search, the question as to whether these tools can be applied to a practical setting such as variant curation and interpretation remains unanswered. (7) For such tools to be useful, they would need to return articles that are relevant to the biocurator's task of classification. Such information includes experimental validation of variant functions, tumour and co-segregation information, family history, *in silico* analysis and statistical methods to determine a probability of pathogenicity.

The literature so far has focussed largely only on the correct identification of gene, mutation and disease within a paper by a literature searching tool. (8) Furthermore, whilst there has been discussion of open source tools applied to breast and prostate cancer variants, analysis of specific applications of literature searching tools with MMR variants is mostly limited to a study which developed 'Variation Annotation Schema' that aimed to capture important concepts and relations for human genetic variation. (9) This schema was developed in response to the needs of InSiGHT biocurators and relates to the historical curation of the InSiGHT database and annotation of MMR genes. It was hoped it would provide a framework for future literature searching tools for MMR variants.

There now exist a range of commercially available literature searching tools and given the onerous task of manual curation, a tool that increases the efficiency or accuracy of the initial literature search could allow the optimum classification of MMR variants and could be beneficial in resolving discordant interpretation. We therefore set out to ask the question as to whether literature searching tools could add incremental value to the initial literature search to retrieve information for the classification of MMR variants submitted with different pathogenicity assignments.

Aims and Hypothesis

Our first aim was to:

- Examine the nature of discordance in a set of MMR variants with different pathogenicity assignments, by comparing InSiGHT classifications to their associated assertions of pathogenicity on the ClinVar database.

Our second aim was to:

- Identify the incremental value that could be added to an initial variant literature Google Scholar search that informs the MMR variant classification process by using the literature searching tool Mastermind Genomenon.

We hypothesised that amongst a sample of variants submitted to ClinVar with different pathogenicity assignments, that Mastermind Genomenon would add incremental value to the initial literature search for a variant being classified in the MMR classification process by providing a more relevant initial literature search and retrieve more unique information compared to a standard Google Scholar search for a particular variant.

Given the importance of Lynch Syndrome as the most common aetiology of hereditary gastrointestinal cancer, if literature searching tools could ease the burden on biocurators, then perhaps the promise of precision medicine could be more easily delivered by the more accurate classification of variants, and the resolution of discordant variants, which would ameliorate some of the associated clinical challenges and risks.

2. Methods

Sample

In January 2020, a list of MMR variants with discordant classifications that require reviewing/updating on the ClinGen ClinVar database was provided to the InSiGHT VCEP by ClinGen as a part of InSiGHT's role in reclassifying these variants. This list was 'prioritised' by ClinGen into several categories. The first of these were the 'Alert' categories: which spanned variants that have existing InSiGHT classifications but now have more recent, differing classifications submitted to ClinVar from other non-expert entities such as laboratories, familial cancer clinics or research institutions. Variants listed as "Priority" are

those variants which do not necessarily have an InSiGHT classification, but now have newer, conflicting classifications on ClinVar from multiple submitters.

To describe the nature of discordance amongst this sample of variants (the first aim) and eventually use the sample to identify the value of Mastermind in an initial variant literature search (the second aim), it was important the sample reflected a typical situation that a biocurator may be faced with, that is: fulfilling the important role of the InSiGHT VCEP by reclassifying VUSs and resolving discordant variants. In order to address the two aims of this study, we used judgement sampling to identify which variants should be selected for inclusion in the study and this was on the basis of priority as designated by ClinVar and number of conflicting submissions on ClinVar. Judgement sampling refers to a sample chosen based on the prior knowledge of a subject and is useful for samples where the aim is to improve process performances, which in our case is the process of literature searching for MMR variant classification. (10)

We first focussed our efforts on the 'Alert' variants and then prioritised a selection of variants from the 'Priority' group. We aimed for an arbitrary total of 80 variants which was thought to be a sufficient enough sample size to pilot the feasibility of Mastermind. As this was intended to be a study that examined the feasibility of using Mastermind across a range of different discordant settings, it was not deemed necessary (and was beyond the scope of this study) to test all of the variants in every category beyond 'Alert'. As detailed statistics were not planned, there was no formal power calculation for sample size.

From the 80 variants on the list, all 31 variants in the 'Alert' category were selected for analysis on the basis of them being of high priority (as designated by ClinGen) for InSiGHT to provide updates on. Further subgroups within the Alert category will be expanded upon in the Results section.

The remaining 49 variants were selected on the basis of multiple submissions with discordant interpretations by different submitters and were from the 'Priority' category as designated by ClinGen. From the 'Priority' category, we focussed on two sub-groups. The first was variants that did not necessarily have an InSiGHT classification but had at least one conflicting pathogenic/likely pathogenic vs VUS/likely benign /benign submission from

different sources on ClinVar, this being a medically significant conflict. To further prioritise these variants, we then derived the median number of submissions to ClinVar per variant, which was four, and selected all the variants with four or more submissions for inclusion. The first of these groups prioritised by this method contained 39 variants.

In the second subgroup of the 'Priority' category, variants did not necessarily have an InSiGHT classification, but had at least one VUS and at least one likely benign/benign non-expert panel classification on ClinVar. Since this group was deemed to be of lower priority and had a large number (540) of variants, we examined the top ten variants with the most assertions of pathogenicity on ClinVar.

Materials/Apparatus/Measures

The Literature Searching Tool

The Mastermind Genomic Search Engine (Mastermind) (https://www.genomenon.com/mastermind) was selected as the commercially available literature searching tool for comparison primarily because of its ease of use (as it does not require the use of Boolean search terms) and popularity amongst biocurators. Mastermind uses artificial intelligence, machine learning and genomic language processing to search the literature for gene variants. To maximise applicability of any results to a general setting we used the Basic, free edition of the software that simply required registration using an email address and password.

The Control

The traditional searching method that we compared the results of Mastermind towas Google Scholar. Google Scholar's ability to search the full text of articles was the primary reason this was used as the standardised control over PubMed, which does not search full text. However, to standardise the Google Scholar search, variants were entered into the Genomizer (www.genomizer.com) interface which parses proteins and variants into the required search terms. For example for the variant c.1984A>C in MLH1 (standardised nomenclature: NM_000249.3(MLH1):c.1984A>C (p.Thr662Pro)) Genomizer parses this to generate a standardised Google Scholar search of : (MLH1 OR NM_000249.3 OR NM_000249) AND (c.1984A>C OR Thr662Pro OR T662P). Such a strategy captures the various ways in which

the variant may be described in literature. The genes and variants were listed along with collected data and observations on an Excel 2016 Spreadsheet.

Procedure / Experimental Protocol

In this study the independent variables were the search methodology (Google Scholar or Mastermind), the dependent variables were the data collected which were: number of articles retrieved for each search method per variant, number of relevant articles, number of articles unique to Mastermind or Google Scholar. The control to which Mastermind results were compared was Google Scholar. Other information collected on each of the variants related to gene information such as the gene name, variant, protein change, InSiGHT classification date, ClinVar submissions of pathogenicity and latest ClinVar submission date.

To address the second aim, whereby the potential value of Mastermind to the initial literature search for a variant was identified, all articles were screened for relevance. Relevance was defined as containing the variant in question as well as data informing the pathogenicity classification such as: tumour information, family history, co-segregation data, *in silico* analysis, functional assays and statistical methods of predicting pathogenicity.

All 80 variants underwent both Google Scholar and Mastermind searches independently of each other. The procedure was as follows: The variant in question was taken from the ClinVar InSiGHT VCEP update list and processed through the Genomizer converter. A Google Scholar search term was generated from Genomizer; patents and citations were excluded from the search results. The number of articles returned was recorded and then subsequently each article was reviewed in full text and the variant mentions in each article scrutinised for relevance according to the definition above. Articles not in English were not counted as relevant articles and duplicates were only accounted for once.

A similar methodology was used for Mastermind whereby each variant was entered into the search interface and the total number of articles returned before and after screening for relevance was recorded. No advanced filters were applied. The articles returned by both methods were then viewed side by side and the number of articles unique to each search method was recorded. Statistical methods planned for this study were descriptive in nature and consisted of frequencies, means, medians and ranges. Further statistical analysis

including testing formally the hypothesis that: Mastermind's results would be more relevant or contain more unique information than Google Scholar's, was deemed beyond the scope of a limited feasibility study that was not randomised nor blinded.

Ethics

Our study met the criteria for a quality assurance study in the Department of Colorectal Medicine and Genetics, The Royal Melbourne Hospital (RMH), Melbourne, Australia. The Office for Research, RMH granted the reference number QA2020043.

3. Results

3.1. The nature of discordance amongst variants selected for inclusion

3.1.1. Genes and corresponding variants selected for inclusion in the study

80 variants across the genes *MLH1, MSH2, MSH6 and PMS2* were examined as described in Table 1. *MLH1* and *MSH2* variants were most common each representing approximately one third of the sample, with *PMS2* variants being the least common.

Table 1: Frequency of variants examined by gene (n=80 variants)

Gene	**Frequency (%)**
MLH1	24 (30.0)
MSH2	25 (31.3)
MSH6	20 (25.0)
PMS2	11 (13.8)

3.1.2. Discordance of InSiGHT assertions of pathogenicity compared to ClinVar assertions of pathogenicity

Table 2 shows the frequency of the different ClinVar assertions of pathogenicity for each of the variants selected, organised by their InSiGHT classification. Whilst InSiGHT provides one classification per variant, ClinVar accepts multiple assertions of pathogenicity from multiple submitters per variant with 357 ClinVar assertions across the 80 variants selected in the study.

Table 2: ClinVar assertions of pathogenicity for the 80 variants organised by InSiGHT classification

	ClinVar Assertions of Pathogenicity		
	ClinVar Pathogenic/Likely Pathogenic	**ClinVar VUS**	**ClinVar Likely Benign/Benign**
InSiGHT Pathogenic/Likely Pathogenic (n= 16/80 variants)	0	32	0
InSiGHT VUS (n= 38/80 variants)	45	108	55
InSiGHT Not classified (n= 26/80 variants)	47	70	0

Of the 80 variants, 16/80 (20%) were classified by InSiGHT as being pathogenic or likely pathogenic. For these 16, there were 32 assertions on the ClinVar database and all were VUS. For the 38/80 (47.5%) of variants that were classified as VUS by InSiGHT 108/208 (51.9%) had ClinVar VUS assertions. However, additionally, 55/208 (26.4%) were ClinVar likely benign/benign and 45/208 (21.6%) were ClinVar pathogenic/likely pathogenic. The remaining 26 variants in the study that were not classified by InSiGHT but had 117 assertions of pathogenicity on the ClinVar database. 47/117 (40.2%) assertions were pathogenic/likely pathogenic and the majority were VUS assertions with 70/117 (59.8%) assertions.

3.2. The incremental value of Mastermind Genomenon in the variant literature search

3.2.1. Total yield and relevance of Google Scholar and Mastermind searches

Table 3 demonstrates that searches in Google Scholar and Mastermind across the 80 variants yielded 477 and 439 articles respectively, giving a total of 916 articles screened for relevance. Per search, Google Scholar on average yielded six articles, compared to Mastermind which on average yielded five articles. However, when screened for relevance, a greater proportion of Mastermind articles (308/429, 70.2%) were deemed relevant when compared to the control, Google Scholar (202/477, 42.3%) Per search, Mastermind yielded more relevant articles on average from the original search when compared to Google Scholar control searches with means of approximately 4 articles and 3 articles respectively.

Table 3: Number of articles yielded by Google Scholar and Mastermind

	Number of articles – GS (control)[1]	**Number of articles – MM[2]**	**Number of relevant articles – GS (control) (% of total)**	**Number of relevant articles – MM (% of total)**
Total	477	439	202 (42.3)	308 (70.2)
- Mean (per search)	5.96	5.49	2.53	3.89
- Median (per search)	3.5	2.5	1.0	2.0
- Range (per search)	0-50	0-63	0-19	0-32
Discordant assertions				
- InSiGHT Pathogenic/Likely Pathogenic vs Newer ClinVar VUS/Likely Benign *(n=16 variants)*	57	34	34 (59.6)	32 (91.3)
- InSiGHT VUS vs Newer ClinVar Pathogenic/Likely Pathogenic */n = 10 variants)*	32	49	12 (37.5)	49 (89.8)
- InSiGHT VUS vs Newer ClinVar Likely Benign/Benign *(n=5 variants)*	21	78	14 (66.7)	15 (19.2)
- ClinVar Pathogenic/Likely vs ClinVar Likely Benign$^{\Psi}$ *(n=39 variants)*	249	173	89 (34.9)	119 (68.8)

- ClinVar VUS vs ClinVar Likely Benign/Benign Φ *(n=10 variants)*	118	93	55 (46.6)	88 (94.6)

[1]Google Scholar, [2]Mastermind, Ψ: Some variants in this category were not necessarily classified by InSiGHT but were in the scope of the InSiGHT Variant Curation Expert Panel (VCEP). They had at least one Pathogenic/Likely Pathogenic ClinVar assertion and at least one VUS/Likely Benign/Benign ClinVar assertion (medically significant conflict). Φ: Variants in this category were not classified by InSiGHT but were in the scope of the InSiGHT VCEP. They had at least one VUS ClinVar assertion and at least one Likely Benign/Benign assertion.

3.2.2. Unique Articles for Google Scholar and Mastermind searches

The number relevant articles that were unique to either Google Scholar or Mastermind can be found in Table 4. Mastermind found an increased proportion of relevant articles that were unique when compared to Google Scholar. (193/308, 62.0% vs 87/202, 43.0%). Additionally, per search, Mastermind had an average of two unique articles, compared to one for Google Scholar. By ClinVar category of discordance, Mastermind returned more unique search results in every category.

Table 4: Unique number of articles across Google Scholar and Mastermind Searches

	Number of Relevant Articles Unique to GS (control) (% of GS relevant articles)	**Number of Relevant Articles Unique to MM (% of MM relevant articles)**
Total	87 (43.0)	193 (62.7)
- Mean (per search)	1.09	2.41
- Median (per search)	0.00	1.00
- Range (per search)	0-7	0-20
Discordant assertions		
- InSiGHT Pathogenic/Likely Pathogenic vs Newer ClinVar VUS/Likely Benign *(n=16 variants)*	19 (55.9)	27 (64.2)
- InSiGHT VUS vs Newer ClinVar Pathogenic/Likely Pathogenic */n = 10 variants)*	4 (33.3)	36 (81.8)
- InSiGHT VUS vs Newer ClinVar Likely Benign/Benign *(n=5 variants)*	5 (35.7)	6 (40.0)
- ClinVar Pathogenic/Likely vs ClinVar New Likely Benign[Ψ] *(n=39 variants)*	36 (41.4)	68 (57.1)
- ClinVar VUS vs Likely Benign/Benign Φ *(n=10 variants)*	23 (41.8)	56 (63.6)

Ψ: as previous. Φ: as previous.

3.2.3. Instances Where Google Scholar or Mastermind returned no information

Table 5 demonstrates that there was a total of 24 instances where either search method returned 0 results. A key finding here is that there were 14 variants for which only one of the two searching methods (7 each) identified articles. Additionally, there were 10 variants for which neither search method found any information.

Table 5: Instances where search methods found no information

Category	**Frequency**
Google Scholar found articles, Mastermind did not	7
Google Scholar did not find articles, Mastermind did	7
Neither Google Scholar nor Mastermind found any articles	10
Total	24

4. Discussion

4.1 The nature of discordance in the sample of MMR variants

The first aim of our study was to describe the nature of discordance amongst variants that were known to be discordant, which would ultimately be used to compare Mastermind to Google Scholar. Of the 80 variants, 16/80 (20%) of variants that were classified by InSiGHT as pathogenic or likely pathogenic now have newer, more recent assertions as a VUS on ClinVar. Additionally, none of the variants in the study had benign/likely benign classifications by InSiGHT, but those that were classified as VUS by InSiGHT did indeed have a substantial number of benign/likely benign classifications on ClinVar, with 55/208 of the assertions being of this nature in ClinVar and another 45 as pathogenic/likely pathogenic. Thus, the most significant features of discordance were newer VUS assertions in the setting of a previously pathogenic InSiGHT classification and the emergence of benign/likely benign classifications by ClinVar submitters that were classified as VUS by InSiGHT. The sample of variants used for determining the utility of literature searching tools in the initial literature search likely reflected a fairly typical setting in which literature searching tools are hypothesised to be of most use, that is, amongst variants with discordant classifications and in particular where newer classifications are VUSs. This could inform future work of the InSiGHT VCEP as it works to resolve discordant classifications and reclassify variants of

unknown significance, notably because such discordant interpretations may have serious clinical consequences.

4.2 The proposed value of Mastermind in a variant literature search

Our second aim was to identify the incremental value that a literature searching tool may add to the initial literature search for the classification of MMR variants. To identify this, we examined overall yield and relevance of Mastermind searches and compared them to Google Scholar results to see whether Mastermind would provide a more relevant initial literature search with a greater proportion of unique information. Whilst Google Scholar results initially returned more articles, after screening each article for relevance, Mastermind returned a greater proportion of articles that were relevant, across most categories of discordance. Whilst variants were not reclassified on the basis of the results of the differing search methods, in terms of incremental value added to the variant classification process, an initial variant search through Mastermind may be more relevant than a traditional Google Scholar search. One could infer that this would allow the biocurator to find more actionable information per search, thereby allowing optimal classification. The biocurator may stop searching after the more efficient Mastermind search, as sufficient evidence might have been gleaned to allow definitive classification.

Another important aspect to consider in the value of literature searching tools is whether they find information not found by traditional methods. To identify this, the relevant articles that were unique to Google Scholar or Mastermind were recorded. In total, 87/202 (40.3%) of relevant Google Scholar articles were unique to Google Scholar, with an average of one unique, relevant article per search. On the other hand, 193/308 (62.0%) of Mastermind relevant articles were unique to Mastermind. Mastermind searches averaged two unique, relevant articles per search. When one considers that the overall average yield for Mastermind before screening for relevance was four articles, this suggests that a substantial proportion of total information found by Mastermind was unique. In terms of the incremental value in the classification process of discordant variants, missing information can be key in resolving classifications. These results suggest that in addition to returning more relevant results, Mastermind was able to add significantly to a Google Scholar search by finding information that would have otherwise not been found. However, 40.3% of relevant Google Scholar articles were also unique – which points to the continued currency of Google Scholar and traditional searching methods.

Another measure useful in examining the potential value of Mastermind was cases where either search method returned no information at all. This is a significant metric if one considers that given the sheer volume of information available, a true zero result may point convincingly to little information for a variant. We sought to see whether Mastermind could find additional information when Google Scholar could not find any information, which might warrant the further use of Mastermind in initial searching. There were 24 variants for which either both or one search method did not return any information. Of these 24, in the case of 10 of them, neither Google Scholar nor Mastermind retrieved any articles – suggesting that very limited data exists for these 10. There were 14 for which only one of the searching strategies (7 each) identified information, which points to Mastermind having a complementary role in the initial literature searching strategy for variant classification.

4.3 Mastermind in the context of previous work on literature searching tools and future directions

To our knowledge, this is the first study that uses a sample of discordant MMR variants and attempts to identify the incremental value that commercial literature searching tools might add to the process of retrieving information related to classification of MMR variants.
Our findings are consistent with the existing opinion on literature searching tools in the MMR space: that literature searching tools, whilst not replacing traditional searching methods, can serve a complementary role in the biocurator's toolkit. (9) Where our study differs from previous work is primarily on the basis of methodology. In our study, Google Scholar tended to return a greater overall search than Mastermind which is contrary to prevailing conclusions in other papers: that automated literature searching tools yield a greater number of publications when compared to traditional search methods. (8) This discrepancy is likely because most papers benchmark to PubMed, which does not search full text, but rather title and abstract. One 2010 study estimated that only 30% of all protein-protein interactions are mentioned in the title and abstract, which PubMed searches are limited by. (11) As such we used Google Scholar which can search full text, which is what Mastermind (and many other emerging literature searching tools) can do. The literature also tends to focus on assessing literature searching tools on the basis of their ability to find natural language paired with mutation mentions i.e the simple occurrence of information within a particular article and

seldom describe the relevance of the information that surrounds the mutation mention.(12) Instead, we sought to assess Mastermind on the basis of screening each article for information relevant to a biocurator reclassifying discordant variants or VUSs. Previous studies have limited their application of literature searching tools to ones that have been designed primarily for research purposes, whereas our study uses a commercial one applied to a specific purpose, that is, literature searching for discordant MMR variant interpretation. (7) In the MMR space, previous work by Verspoor et. al developed a Variant Annotation Schema which was hoped to be the basis of future literature searching tools. (9) This study builds on this work by showing the potential value of a commercially available literature searching tool used in the initial literature search for variants in the setting where information directly related to classification is being sought.

In addressing generalisability, our first aim established that our sample contained a significant number of variants with newer pathogenic/likely pathogenic or likely benign classifications on the ClinVar database than their existing InSiGHT classifications. Being varied in discordance presents a typical setting in which a biocurator may use literature searching tools to conduct an initial literature search to classify discordant variants. (13) An increase in sample of variants would increase generalisability of these results and other searching tools could be trialled to explore the utility of such tools other than Mastermind. Limitations of this study lie in the fact that it was a single investigator study; in the future the likely inter-operator variability of searching could be addressed by deploying more investigators. Further extensions with the use of F-measures and statistical hypothesis testing methods may make this work more comparable to the existing literature. (14) In terms of future directions, one may attempt to address processes beyond the initial literature search to assess whether information found by literature searching tools was later actively used in formal classification of variants by groups such as the InSiGHT VCEP. The current work will usefully inform the work of the InSiGHT VCEP as it works to reach a consensus on the pathogenicity of the discordant variants studied here. Structured interviews with biocurators may be helpful in quantifying their opinions on emerging literature searching tools, as the literature only points to a small survey of 30 biocurators in 2012, which, given the emergence newer commercially available searching tools, may be outdated. (15)

5. Conclusion

Our study has showed that for a sample of MMR variants with discordant classifications, Mastermind added incremental value to the initial literature search for a variant in question by providing a greater proportion of relevant articles overall and on average per search. We identified that Mastermind presented a greater proportion of unique articles not found by a Google Scholar search, highlighting its potential to source information missed by traditional searching methods. Given Mastermind still missed some information it would not completely replace Google Scholar, but would be a very useful, complementary feature in the biocurator's variant interpretation toolkit.

Optimal MMR variant classification relies on the biocurator not only being able to retrieve a comprehensive literature search but also accessing information identified as relevant to the purpose of variant classification. The literature search should also not miss key information that might hold important answers related to optimal classification. Being an onerous task, if literature searching tools are able to add value to the initial search process and hence the overall classification process, then one may ultimately be able to resolve discordant interpretations and reclassify VUSs more efficiently and more accurately. The InSiGHT VCEP is committed to this task and delivering on the promise of precision medicine for patients and their families where it is hoped that literature searching tools may play a valuable role in this effort.

Word count: 4393 words

List of Abbreviations

MMR: Mismatch repair
LS: Lynch Syndrome
InSiGHT: International Society for Gastrointestinal Hereditary Tumours
VUS: Variant of unknown significance
NHGRI: US National Human Genome Research Institute's (NHGRI)
NCBI: US National Centre for Biotechnology Information (NCBI)

Acknowledgements

This work was possible due to the efforts of Professor Finlay A. Macrae and John-Paul Plazzer on behalf of the Royal Melbourne Hospital and InSiGHT.

References

1. Biller LH, Syngal S, Yurgelun MB. Recent advances in Lynch syndrome. Familial Cancer. 2019;18(2):211-219.
2. Win AK, Jenkins MA, Dowty JG, Antoniou AC, Lee A, Giles GG, et al. Prevalence and Penetrance of Major Genes and Polygenes for Colorectal Cancer. Cancer Epidemiology Biomarkers & Prevention. 2017;26(3):404-412.
3. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen — The Clinical Genome Resource. New England Journal of Medicine. 2015;372(23):2235-42.
4. Sijmons RH, Greenblatt MS, Genuardi M. Gene variants of unknown clinical significance in Lynch syndrome. An introduction for clinicians. Familial Cancer. 2013;12(2):181-187.
5. Murray ML, Cerrato F, Bennett RL, Jarvik GP. Follow-up of carriers of BRCA1 and BRCA2 variants of unknown significance: Variant reclassification and surgical decisions. Genetics in Medicine. 2011;13(12):998-1005.
6. Allot A, Peng Y, Wei CH, Lee K, Phan L, Lu Z. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic Acids Res. 2018;46(W1):W530-W536.
7. Wei CH, Harris BR, Li D, Berardini TZ, Huala E, Kao HY, et al. Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. Database (Oxford). 2012;2012:bas041.
8. Wei CH, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. Bioinformatics. 2018;34(1):80-87.
9. Verspoor K, Jimeno Yepes A, Cavedon L, McIntosh T, Herten-Crabb A, Thomas Z, et al. Annotating the biomedical literature for the human variome. Database : The Journal of Biological Databases and Curation. 2013;2013:bat019.
10. Perla RJ, Provost LP. Judgment Sampling: A Health Care Improvement Perspective. Quality Management in Healthcare. 2012;21(3):169-175.
11. Harmston N, Filsell W, Stumpf MP. What the papers say: text mining for genomics and systems biology. Hum Genomics. 2010;5(1):17-29.
12. Singhal A, Simmons M, Lu Z. Text mining for precision medicine: automating disease-mutation relationship extraction from biomedical literature. Journal of the American Medical Informatics Association : JAMIA. 2016;23(4):766-772.

13. Harrison SM, Dolinsky JS, Knight Johnson AE, Pesaran T, Azzariti DR, Bale S, et al. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. Genet Med. 2017;19(10):1096-1104.
14. Verspoor KM, Heo GE, Kang KY, Song M. Establishing a baseline for literature mining human genetic variants and their relationships to disease cohorts. BMC Medical Informatics and Decision Making. 2016;16 Suppl 1:68.
15. Hirschman L, Burns GAPC, Krallinger M, Arighi C, Cohen KB, Valencia A, et al. Text mining for the biocuration workflow. Database : The Journal of Biological databases and Curation. 2012;2012:bas020