JHIA - online.org
Journal of Health Informatics in Africa

Diabetes information retrieval research

Kyle Andrew Fitzgerald [a], Jane Alice Fitzgerald [a], Andrew John Bytheway [b]

[a] Cape Peninsula University of Technology, Cape Town, South Africa
[b] University of the Western Cape, Bellville, South Africa

Background and Purpose: For researchers and others making use of information retrieval systems, choosing the most effective phrase terms to retrieve relevant documents remains a challenge. The purpose of this study is to establish, test and evaluate a standard set of phrase terms within a text collection.
Methods: 52 phrase terms were extrapolated from the literature concerning diabetes, and were used to create nine queries each relating to a diabetes classification. A specificity information retrieval system was used to assess and retrieve documents using those queries. Results were analysed to measure research interest and phrase term usage.
Results: 9,106 documents were retrieved from the collection. Diabetes research interest is in: 'type 2 diabetes mellitus', 'type 1 diabetes mellitus' and 'gestational diabetes mellitus' with the classification 'type 2 diabetes mellitus' having three times more research interest than 'type 1 diabetes mellitus'. The top five frequently used phrase terms were: 'type 2 diabetes', 'type 1 diabetes', 'diabetes mellitus', 'type 2 diabetes mellitus' and 'prediabetes'.
Conclusions: Most research interest is vested in 'type 2 diabetes mellitus' and 'type 1 diabetes mellitus'. Research interest is increasing for 'prediabetes' and 'gestational diabetes mellitus'. Phrase term usage tends to increase when research interest is low.

Keywords: Information retrieval, query, phrase term usage, diabetes, research interest.

1 Introduction

The general challenges in retrieving special-interest documents have prompted experts to test and evaluate different ideas and theories of information retrieval [4] [5] [6]. These evaluations apply to many disciplines [4] [5] but in particular to healthcare [6]. The goal of a retrieval system is to improve the efficiency of document retrieval from a collection and to retrieve those documents that are most relevant to a user's information need [4] [5], but for academics, researchers, family members and others involved in healthcare [1] making use of the Web [2], electronic documents [3] and various information retrieval systems [4] to find relevant documents about diabetes remains a particular challenge.

Various multi-word phrase terms are typically used to retrieve documents that might be relevant [5] to a particular classification of diabetes but the phrase terms used to describe a diabetes classification have evolved over time [7]. There are two main forms of diabetes: references to one have evolved from: 'fat diabetes' [8], to 'adult onset diabetes' [9], to 'non insulin dependent diabetes mellitus' [10], to 'type 2 diabetes mellitus' [7]; references to the other have evolved from: 'thin diabetes' [8], to 'juvenile diabetes' [11], to 'insulin dependent diabetes mellitus' [12], to 'type 1 diabetes mellitus' [8]). Hence using an information retrieval system to search for these phrase terms, measuring the usage of these phrase terms, and measuring the research interest into diabetes classifications becomes challenging. The queries used in the searches must be carefully constructed, making use of appropriate phrase terms to improve the efficiency of retrieval and the relevance of the documents retrieved [5].

*Corresponding author

J Health Inform Afr. 2017;4(2):xx-xx. 10.12856/JHIA-2017-v4-i2-152

In this exploratory study, existing literature was collected and analysed in order to understand and document the phrase terms used to identify the different diabetes classifications, and to reveal the frequency of interest and usage.

2 Purpose of the study

The overall objective is to assist academics, researchers, family members and others to understand how diabetes classifications are structured, and to develop terms to use in a search query. With this in mind, the particular purpose of this study was to examine the use of a custom-designed specificity information retrieval system employing two indexes with multi-word phrase term search capabilities; one index based on *the content of the conference texts*, and the other on *the content of the queries* that might be used to search them.

Two research questions are addressed:

- What is the level of research interest in the various diabetes classifications, based on the occurrence of the standard phrase terms?
- What phrase terms are most frequently used to describe a form of diabetes in some way?

With this understanding, it becomes possible to re-organise the way that specialist literature such as this is accessed, and to bring together the various vocabularies and ontologies that specialists use to render their work comprehensible and meaningful as well as accessible.

3 Materials and methods

The published material that was used to exercise the specificity information retrieval system was drawn from five conferences of the International Diabetes Federation (IDF). 9,106 IDF conference paper abstracts and posters were downloaded from the five IDF conferences held over the past ten years: Cape Town (2006), Montreal (2009), Dubai (2011), Sydney (2013) and Vancouver (2015). First, a standard set of phrase terms was derived from the collected texts; then, the occurrence of those phrase terms was analysed across all the articles within the collection.

A custom-designed information retrieval system was used to process phrase term queries efficiently. The original design approach was design science research [40] [41] [42] [43] that led to a trio of new artefacts: the two hybrid indexes and the specificity information retrieval system that employed them. The specificity information retrieval system has a dual process (Figure 1): the first gathers information from the documents in the collection, and the second processes queries using its search engine [5] [4]. All data pertaining to the documents is placed in the information retrieval systems data store [4].

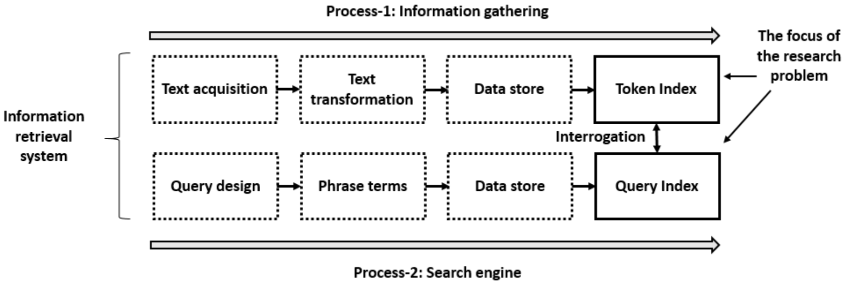

Figure 1: The building blocks of the specificity information retrieval system

3.1 Developing the token index

The first stage, information gathering, used the specificity information retrieval system to acquire, transform and store the textual content in a hybrid token index. All documents in each of the five sub-collections were converted to text files, case folded to lowercase [5], with white spaces and special characters replaced with a pipe delimiter [44], thereby allowing words (or tokens) to be extracted from the texts; these tokens were then used to populate a *hybrid token index*, keeping word ordinality and proximity intact. In essence, reading the contents of the hybrid token index (represented as a database table) from top-to-bottom mirrored a reading of the original text from left-to-right.

3.2 Developing the phrase index

With the data collection and organisation now complete, the second stage (the search engine processes) could begin. This stage is described in some detail, as the provenance of the phrases to be included in the index is important. 52 phrase terms extrapolated from the healthcare literature (each describing a variety of diabetes in some way) were arranged into nine diabetes classifications hierarchically using the four levels illustrated in Figure 2. The sources of the terms are provided following the explanation of the figure.

The top node of the hierarchy represents classification-1 at level-1. Diabetes itself is not a singular disease but a group of diseases [7] that at level-2 is divided into two classifications: classification-2 as ‘diabetes mellitus’ and classification-3 as ‘diabetes insipidus’ with the former comprising of multiple different forms. Therefore ‘diabetes mellitus’ is a combination of at least four disparate forms of diabetes at level-3: these are described as classification-4 for ‘gestational diabetes mellitus’, classification-5 for ‘type 1 diabetes mellitus’, classification-6 for ‘type 2 diabetes mellitus’ and classification-7 for what we call ‘other forms of diabetes mellitus’ - a collection of phrase terms describing other forms of diabetes not officially classified as a type. To understand which phrase terms to use in a search for each of these four classifications at level-3 we need to gather the information, those phrase terms authors have used historically and currently, from the literature.

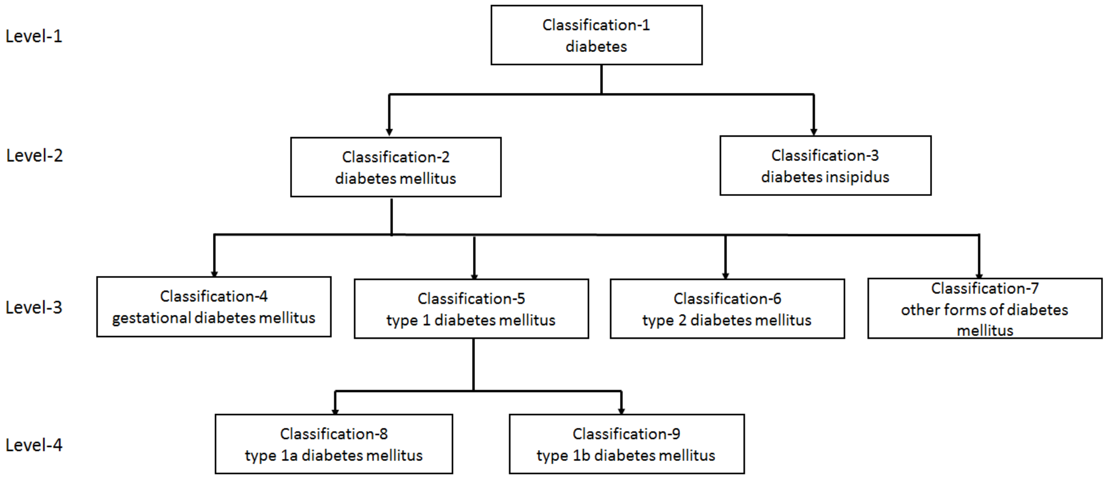

Figure 2: A hierarchy of diabetes classifications

Firstly, we begin with ‘gestational diabetes mellitus’ where two phrase terms were retrieved from the literature ‘gestational diabetes mellitus’ [13] and ‘gestational diabetes’ without the mellitus [14].

Secondly, we have ‘type 1 diabetes mellitus’ that is the only classification at level-3 that has a lower level. Therefore at level-4 there are two sub-types of ‘type 1 diabetes mellitus’ and these are ‘type 1a diabetes mellitus’ related to an autoimmune process [7] and ‘type 1b diabetes mellitus’ [7] related to an unknown cause. The phrase terms used to describe ‘type 1a diabetes mellitus’ from the literature are: 'autoimmune diabetes' [15], 'autoimmune type 1 diabetes' [16], 'immune mediated type 1 diabetes' [17] and 'type 1a diabetes' [7]. The phrase terms used to describe ‘type 1b diabetes mellitus’ from the literature are: 'idiopathic diabetes' [8], ‘idiopathic type 1 diabetes' [18] and 'type 1b diabetes' [7]. Other phrase terms not specific to the two sub-types are: 'brittle diabetes' [19], 'diabetes mellitus type 1' [20], 'diabetes type 1' [21], 'insulin dependent diabetes' [22], 'insulin dependent diabetes mellitus' [12], 'juvenile diabetes' [11], 'juvenile onset diabetes' [22], 'juvenile onset type diabetes' [22], 'slowly progressive insulin dependent diabetes mellitus' [23], 'spontaneous autoimmune diabetes' [24], 'thin diabetes' [8], 'type 1 diabetes' [25], 'type 1 diabetes mellitus' [26], 'type i diabetes' [27] and 'type i diabetes mellitus' [16].

Thirdly, we have ‘type 2 diabetes mellitus’ with 11 phrase terms retrieved from the literature and these are: 'adult onset diabetes' [9], 'diabetes mellitus type 2' [28], 'diabetes type 2' [29], 'fat diabetes' [8], 'non insulin dependent diabetes' [22], 'non insulin dependent diabetes mellitus' [10], 'stable diabetes' [30], 'type 2 diabetes' [31], 'type 2 diabetes mellitus' [32], 'type ii diabetes' [33] and 'type ii diabetes mellitus’ [33].

Fourth and finally, we have ‘other forms of diabetes mellitus’ with 16 phrase terms retrieved from the literature and these are: 'ketosis prone diabetes' [22], 'ketosis resistant diabetes' [34], 'latent autoimmune diabetes in adults' [35], 'latent diabetes' [36], 'malnutrition modulated diabetes' [23], 'malnutrition modulated diabetes mellitus' [37], 'malnutrition related diabetes' [38], 'malnutrition related diabetes mellitus' [37], 'maturity onset diabetes' [8], 'maturity onset diabetes of the young' [31], 'maturity onset type diabetes' [22], 'potential diabetes' [22], 'prediabetes' [39], 'protein deficient diabetes mellitus' [37] and ‘'secondary diabetes' [36].

3.3 Design of queries

With the hierarchy of phrases, a classification can now be represented by a query containing one or more of these phrase terms. For the top node at classification-1 we use the anchor word *diabetes* as it occurs in each of the 52 phrase terms, surrounded by inverted commas, and is represented by query 1 as: q01 = [“diabetes”]. For classifications two through to nine we make use of query expansion [4] where each phrase term is sequentially placed and separated by the Boolean logical OR operator [45]. For example, the query for classification-8 at level-4 ‘type 1a diabetes mellitus' is represented as: q08 = [“autoimmune diabetes” OR “autoimmune type 1 diabetes” OR “immune mediated type 1 diabetes” OR “type 1a diabetes”].

To exercise information retrieval, each of the nine queries was submitted to the search engine, thus creating the hybrid query index “on the fly”; those documents judged by the information retrieval system as ‘relevant’ and those that were judged ‘non relevant’ [5] were retrieved and the data store [4] was populated with the document statistics ready for data analysis.

3.4 Data analysis

Data analysis was performed utilising the information retrieval systems data store making use of the measurements: document frequency and collection frequency [5] [4]. ***Research interest*** is established as the number of documents (retrieved as “relevant”) that contain at least one of the phrase terms in the query. ***Phrase term usage*** is the number of times a phrase term occurs, here measured within each of the five collections separately and together. During analysis we allowed for phrase term co-existence [46], a phenomenon that occurs when one phrase term co-exists within another, for example, *‘diabetes mellitus’* and ‘*type 1 diabetes mellitus’* where the former co-exists within the latter.

4 Results

The results are presented in Tables 1, 2 and 3; the results are discussed in the section that then follows. First, it is useful to summarise the number of documents in each sub-collection from the five IDF conferences. Table-1 below presents the results for these five sub-collections and the total number of documents in the collection.

Table 1. Number of documents in collection

IDF Year	IDF conference venue	No of documents in collection, N
2006	Cape Town	2,123
2009	Montreal	1,739
2011	Dubai	1,833
2013	Sydney	1,891
2015	Vancouver	1,520
Total		9,106

4.1 Research question 1 - Research interest

The first research question was to determine the research interest into the various types of diabetes classifications. Table-2 presents the results of research interest per diabetes classification in rank order using the descending document frequency and as a percentage of the document collection.

Table 2. Research interest

Rank	Classification	qt	2006		2009		2011		2013		2015		Total	
			df	%	df	%	df	%	df	%	df	%	df	%
1	diabetes	q01	2,123	100	1,556	89.48	1,648	89.91	1,711	90.48	1,387	91.25	8,426	92.53
2	diabetes mellitus	q02	1,213	57.14	981	56.42	1,042	56.85	1,209	63.93	939	61.78	5,384	59.13
3	type 2 diabetes mellitus	q06	843	39.71	682	39.22	704	38.41	809	42.78	606	39.87	3,644	40.02
4	type 1 diabetes mellitus	q05	282	13.28	189	10.87	205	11.18	264	13.96	199	13.09	1,139	12.51
5	gestational diabetes mellitus	q04	55	2.59	59	3.39	70	3.82	79	4.18	70	4.61	333	3.66
6	other forms of diabetes mellitus	q07	25	1.18	47	2.7	55	3	52	2.75	55	3.62	234	2.57
7	type 1a diabetes mellitus	q08	8	0.38	4	0.23	7	0.38	5	0.26	9	0.59	33	0.36
8	diabetes insipidus	q03	0	0	0	0	0	0	1	0.05	0	0	1	0.01
8	type 1b diabetes mellitus	q09	0	0	0	0	1	0.05	0	0	0	0	1	0.01

4.2 Research question 2 – Phrase term usage

The second research question was to determine the usage of phrase terms used to describe a form of diabetes in some way. Table-3 presents the results of phrase term usage in rank order using the descending collection frequency over the ten-year period.

Table 3. Phrase term usage

Rank	pt	Phrase term	cf 2006	cf 2009	cf 2011	cf 2013	cf 2015	cf Total
1	pt47	type 2 diabetes	1,483	1,226	1,176	1,376	954	6,215
2	pt43	type 1 diabetes	514	312	321	481	347	1,975
3	pt06	diabetes mellitus	405	349	394	223	162	1,533
4	pt48	type 2 diabetes mellitus	234	211	233	241	194	1,113
5	pt36	Prediabetes	19	84	138	117	109	467
6	pt12	gestational diabetes	65	99	81	73	88	406
7	pt13	gestational diabetes mellitus	43	28	57	71	52	251
8	pt44	type 1 diabetes mellitus	66	41	44	43	32	226
9	pt10	diabetes type 2	30	19	23	11	12	95
10	pt08	diabetes mellitus type 2	24	20	14	7	9	74
11	pt09	diabetes type 1	26	6	14	13	4	63
12	pt02	autoimmune diabetes	8	10	9	9	12	48
13	pt51	type ii diabetes	19	0	8	10	8	45
14	pt07	diabetes mellitus type 1	11	8	11	3	1	34
15	pt49	type i diabetes	10	4	9	5	0	28

16	pt19	juvenile diabetes	2	2	3	3	11	21
17	pt31	maturity onset diabetes of the young	5	7	0	5	2	19
18	pt17	insulin dependent diabetes	7	1	3	3	4	18
19	pt38	secondary diabetes	5	7	3	0	2	17
20	pt24	latent autoimmune diabetes in adults	7	2	1	2	1	13
21	pt33	non insulin dependent diabetes	4	1	5	0	0	10
22	pt03	autoimmune type 1 diabetes	1	1	1	2	4	9
22	pt18	insulin dependent diabetes mellitus	4	1	0	3	1	9
22	pt52	type ii diabetes mellitus	0	0	4	4	1	9
23	pt01	adult onset diabetes	1	1	0	2	3	7
24	pt22	ketosis prone diabetes	0	1	0	1	3	5
25	pt29	malnutrition related diabetes mellitus	2	0	0	2	0	4
25	pt35	potential diabetes	1	1	2	0	0	4
25	pt45	type 1a diabetes	3	1	0	0	0	4
26	pt04	brittle diabetes	2	0	0	1	0	3
26	pt50	type i diabetes mellitus	0	1	1	1	0	3
27	pt34	non insulin dependent diabetes mellitus	0	1	0	1	0	2
28	pt05	diabetes insipidus	0	0	0	1	0	1
28	pt14	idiopathic diabetes	0	0	1	0	0	1
28	pt20	juvenile onset diabetes	0	0	0	1	0	1
28	pt27	malnutrition modulated diabetes mellitus	0	0	0	0	1	1
28	pt37	protein deficient diabetes mellitus	1	0	0	0	0	1
28	pt39	slowly progressive insulin dependent diabetes mellitus	1	0	0	0	0	1
28	pt40	spontaneous autoimmune diabetes	0	0	0	0	1	1
29	pt11	fat diabetes	0	0	0	0	0	0
29	pt15	idiopathic type 1 diabetes	0	0	0	0	0	0
29	pt16	immune mediated type 1 diabetes	0	0	0	0	0	0
29	pt21	juvenile onset type diabetes	0	0	0	0	0	0
29	pt23	ketosis resistant diabetes	0	0	0	0	0	0
29	pt25	latent diabetes	0	0	0	0	0	0
29	pt26	malnutrition modulated diabetes	0	0	0	0	0	0
29	pt28	malnutrition related diabetes	0	0	0	0	0	0
29	pt30	maturity onset diabetes	0	0	0	0	0	0
29	pt32	maturity onset type diabetes	0	0	0	0	0	0
29	pt41	stable diabetes	0	0	0	0	0	0
29	pt42	thin diabetes	0	0	0	0	0	0
29	pt46	type 1b diabetes	0	0	0	0	0	0

5 Discussion

5.1 Number of documents in the collections

Table-1 shows that the total number of documents in the collection, from the five IDF conferences, was 9,106. The number of documents per sub-collection varied over the years with 2,124, 1,729, 1,833, 1,891 and 1,520 representing Cape Town (2005), Montreal (2009), Dubai (2011), Sydney (2013) and Vancouver (2015) respectively. The sub-collections continually declined in volume from a peak at the first conference in Cape Town; the conference in Sydney showed a slight improvement in volume over the previous conference held in Dubai, but these variations are seen as spurious because of the large number of unknown factors involved.

5.2 Research question 1 – Research interest

The first research question concerned the level of research interest in the various types of diabetes classifications. The results from Table-2 are now discussed in rank order using descending document frequency as a percentage of the document collection.

Rank-1 - Diabetes
Of the nine classifications, 'diabetes' was ranked first, with the highest document frequency thus representing the highest research interest. Referring to Table-2, and for Classification-1 at Level-1, a total of 92.53% or 8,426 of 9,106 documents were relevant to diabetes in some form as all these documents contained at least one occurrence of the phrase term 'diabetes' in their text. Conversely 9,106 – 8,426 = 680 documents did not contain the phrase term 'diabetes' and therefore did not refer to this classification directly. In 2006 all documents referred to diabetes but over the following four conferences they consistently averaged around 90% suggesting there were other areas of research interest bundled into these conferences.

Rank-2 - Diabetes mellitus
Ranked second was 'diabetes mellitus'. The results would have been based on a combination of unique occurrences of the phrase term (phrase term co-existence was considered) not related to the other phrase terms in the queries in addition to the 50 other phrase terms (phrase terms 'diabetes' and 'diabetes insipidus' were excluded from this query). For this classification-2 at Level-2, a total of 59.13% or 5,384 of 9,106 documents were relevant to 'diabetes mellitus' in some form as all these documents contained at least one occurrence of the phrase term 'diabetes mellitus' or at least one other phrase term within query number two. During the first three conferences, the research interest purely into this classification averaged 57% but this increased to around 62% for the latter two.

Rank-3 – Type 2 diabetes mellitus
Ranked third was 'type 2 diabetes mellitus' and for this classification-6 at Level-3, a total of 40.02% or 3,644 of 9,106 documents were relevant in some form as all these documents contained at least one occurrence of the eleven phrase terms in its query. During four of the conferences the research interest into this classification averaged 40% but this increased to nearly 43% in 2013 suggesting either more research interest into 'type 2 diabetes mellitus' or less interest into the other classifications.

Rank-4 – Type 1 diabetes mellitus
Ranked fourth was 'type 1 diabetes mellitus' and for this classification-5 at Level-3, a total of 12.51% or 1,139 of 9,106 documents were relevant in some form. During 2009 and 2011 the research interest into this classification dropped below the average while in other years was just above average.

Rank-5 – Gestational diabetes mellitus
Ranked fifth was 'gestational diabetes mellitus' the form that occurs in pregnant woman. For this classification-4 at Level-3, a total of 3.66% or 333 of 9,106 documents were relevant in some form to 'gestational diabetes mellitus'. The research interest into this classification has continually increased over the ten year period with 2.59%, 3.39% 3.82%, 4.18% and 4.61% possibly suggesting a new focus for diabetes research.

Rank-6 – Other forms of diabetes mellitus
Ranked sixth was 'other forms of diabetes mellitus' where specific diabetes typing for these forms has not occurred. For this classification-7 at Level-3, a total of 2.57% or 234 of 9,106 documents were relevant in some form. Similar to 'gestational diabetes mellitus' the research interest into this classification has continually increased over the ten year period with the exception of 2013. The percentages of interest were 2.5%, 2.7%, 3%, 2.75% and 3.62% for 2006, 2009, 2011, 2013 and 2015 respectively. These figures suggest research interest for this classification is growing, albeit slowly.

Rank-7 – Type 1a diabetes mellitus
Ranked seventh was 'type 1a diabetes mellitus' one of two sub-classifications for 'type 1 diabetes mellitus'. For this classification-8 at Level-4, a total of 0.36% or 33 of 9,106 documents were relevant in some form.

During 2009 and 2013 the research interest into this classification dropped below the average while in the other three years they continued to increase from 0.38% in 2006 to 0.59% in 2015. This form of diabetes is the deadly one that traditionally targets young children and creates an autoimmune response. These figures suggest that at last there is a slight increase in research interest into this classification.

Rank-8 – Diabetes insipidus and type 1b diabetes mellitus
Ranked eight were both 'diabetes insipidus' and 'type 1b diabetes mellitus'. The former is the only non 'diabetes mellitus' form of diabetes in this research. For this classification-3 at Level-2 only one document was retrieved relevant to this classification suggesting very low, if not insignificant, research interest. A similar result was obtained for 'type 1b diabetes mellitus' again with only one document over the ten years. This form of diabetes is also a deadly one that traditionally targets young children but does not create an autoimmune response.

5.3 Research question 2 – Phrase term usage

The second research question concerned the usage of phrase terms used to describe a form of diabetes in some way. The results from Table-3 are now discussed in rank order of descending collection frequency.

No usage – Rank 29
Of the 9,106 documents in the collection the usage of 13 of the 52 phrase terms (rank 29) were not evidenced at all. These were: 'fat diabetes', 'idiopathic type 1 diabetes', 'immune mediated type 1 diabetes', 'juvenile onset type diabetes', 'ketosis resistant diabetes', 'latent diabetes', 'malnutrition modulated diabetes', 'malnutrition related diabetes', 'maturity onset diabetes', 'maturity onset type 'diabetes', 'stable diabetes', 'thin diabetes' and 'type 1b diabetes'. Interestingly even though 'type 1b diabetes' is a type of diabetes its alternative synonymic phrase term 'idiopathic diabetes' was the preferred choice albeit only once in the collection.

Low usage – Rank 22 to 28
From the document collection 18 of the 52 phrase terms (rank 22 to 28) had low usage where their collection frequencies were below ten. In rank order these were: 'autoimmune type 1 diabetes', 'insulin dependent diabetes mellitus', 'type ii diabetes mellitus', 'adult onset diabetes', 'ketosis prone diabetes', 'malnutrition related diabetes mellitus', 'potential diabetes', 'type 1a diabetes', 'brittle diabetes', 'type i diabetes mellitus', 'non insulin dependent diabetes mellitus'. Ranked 28th with a collection frequency of 1 were" 'diabetes insipidus' the only non diabetes mellitus related phrase term and 'idiopathic diabetes' a phrase term used to describe a form of 'type 1b diabetes mellitus'. Then 'juvenile onset diabetes' similar to 'juvenile diabetes' and an antonym to 'adult onset diabetes' once popular phrase terms to describe 'type 1 diabetes mellitus' and 'type 2 diabetes mellitus' respectively. Other phrase terms with a low usage collection frequency of 1 were: 'malnutrition modulated diabetes mellitus', 'protein deficient diabetes mellitus', 'slowly progressive insulin dependent diabetes mellitus' and 'spontaneous autoimmune diabetes'. Ranked 22nd with nine occurrences were: 'autoimmune type 1 diabetes', 'insulin dependent diabetes mellitus', 'type ii diabetes mellitus', followed by 'adult onset diabetes' with seven,, 'ketosis prone diabetes' with five, 'malnutrition related diabetes mellitus', 'potential diabetes' and 'type 1a diabetes' all with four, 'brittle diabetes', 'type i diabetes mellitus', with three and finally 'non insulin dependent diabetes mellitus' with two occurrences. The phrase term 'non insulin dependent diabetes mellitus' once a popular synonym for 'type 2 diabetes mellitus' has reduced in usage over the years. Those low usage phrase terms that have increased in usage over the years are: 'autoimmune type 1 diabetes', 'adult onset diabetes' and 'ketosis prone diabetes'.

Medium usage – Rank 9 to 21
From the document collection 13 of the 52 phrase terms (rank 9 to 21) had medium usage where their collection frequencies of between 95 and 10. In rank order these were: 'diabetes type 2', 'diabetes mellitus type 2', 'diabetes type 1', 'autoimmune diabetes', 'type ii diabetes', 'diabetes mellitus type 1', 'type i diabetes', 'juvenile diabetes', 'maturity onset diabetes of the young', 'insulin dependent diabetes', 'secondary diabetes', 'latent autoimmune diabetes in adults', and 'non insulin dependent diabetes'. It is interesting to evident word order reversal in a number of the phrase terms and the use of letters as roman numerals instead of numbers for example: 'diabetes mellitus type 2' and 'type ii diabetes'. This supports the design of the specify

information retrieval system and its pair of indexes that keep word ordinality and proximity intact. The usage of most of these phrase terms are in decline except for: ‘autoimmune diabetes’ increasing in usage from 8 to 12 in 2006 and 2015 respectively and ‘juvenile diabetes’ increasing from 2 to 11 in 2006 and 2015 respectively. Again, the phrase term ‘non insulin dependent diabetes’, without the ‘mellitus’, has reduced in usage over the years to zero in 2015.

High usage – Rank 1 to 8

The high usage phrase terms evidenced for the document collection ranked 1 to 8 encompass eight phrase terms with collection frequencies of between 6,215 and 226 and these are: 'type 2 diabetes', 'type 1 diabetes', 'diabetes mellitus', 'type 2 diabetes mellitus', 'prediabetes', 'gestational diabetes', 'gestational diabetes mellitus' and 'type 1 diabetes mellitus'. These eight phrase terms represent a formal classification in the diabetes hierarchy with exceptions of using or not using the word ‘mellitus’. Ranked first, the most used phrase term to describe a diabetes classification in some form was 'type 2 diabetes' without the ’mellitus’ has decreased in usage from 1,483 occurrences in 2006 to 954 in 2015. Similarly, with a rank of 4, ‘type 2 diabetes mellitus' with the ’mellitus’ has decreased in usage from 234 occurrences in 2006 to 194 in 2015. With a rank of 2 ‘type 1 diabetes’ without the ‘mellitus’ decreased in usage from 514 to 347 and ranked 8 ‘type 1 diabetes mellitus’ with the ‘mellitus’ decreased in usage from 66 to 32 over the ten year period. The usage of the phrase term ‘diabetes mellitus’, that excludes its use within other phrase terms because phrase term co-existence was excluded, was ranked 3 but also decreased in usage. Those phrase terms that did increase in usage were 'prediabetes', 'gestational diabetes' and 'gestational diabetes mellitus' ranked 5, 6 and 7 respectively. 'prediabetes' increased from 19 to 109 occurrences while 'gestational diabetes' and 'gestational diabetes mellitus' increased from 65 and 43 to 88 and 52 respectively.

6 Conclusion

6.1 Frequencies and trends

The five most frequently used phrase terms used to describe a form of diabetes in some way were: ‘type 2 diabetes’, ‘type 1 diabetes’, ‘diabetes mellitus’, ‘type 2 diabetes mellitus’ and ‘prediabetes’. Research interest into ‘type 2 diabetes mellitus’ is more than three times that of ‘type 1 diabetes mellitus’. Within the classification ‘type 1a diabetes mellitus’ the phrase term ‘type 1a diabetes’ is one of the least used with a collection frequency of 4. The most popular are: ‘autoimmune diabetes’, ‘insulin dependent diabetes’, ‘autoimmune type 1 diabetes’ and ‘insulin dependent diabetes mellitus’. Within the classification ‘type 1b diabetes mellitus’ the phrase term ‘type 1b diabetes’ was never used, only the phrase term ‘idiopathic diabetes’ was used and only once. There is a tendency for authors to drop the mellitus when describing a diabetes classification in some form. Phrase term usage tends to increase when research interest into a specific diabetes classification is low. Research interest into ‘prediabetes’ and ‘gestational diabetes mellitus’ is increasing albeit slowly.

6.2 Successful retrieval using queries

What was achieved in this study was the successful and efficient retrieval of documents relevant to each of the nine classifications of diabetes. Through the use of expanded queries making use of specific phrase terms that have described a diabetes classification in some way over time, all documents relevant to each diabetes classification were retrieved judged relevant by the information retrieval system. By defining the phrase terms to use upfront, and by using query expansion to increase the size of the net, more relevant documents were retrieved (and fewer that were not relevant).

6.3 Outcome

The 52 phrase terms that were developed are found to describe diabetes classifications effectively for the purposes of retrieval. There may be many more phrase terms used in the literature, but this is a good starting point to help experts and lay-people to search the literature and to retrieve documents that are more relevant, more easily. As research across the world evolves and becomes more global, and as informed patients and carers read more deeply into such areas of special interest, it is important that the words and phrases that are used are properly defined and understood. The work reported here is a first step towards a future where it will be possible to improve the way that specialist literature is organised and accessed, and to bring together the work of experts in a more comprehensible and meaningful way.

Acknowledgements

We acknowledge the IDF-copyrighted material conditions for the reproduction and translation of IDF publications and that the source material containing the abstracts and posters residing on compact disc and the Web for the years of 2006, 2009, 2011, 2013 and 2015 was supplied by the IDF.

Statement on conflicts of interest

There was no conflict of interest in this study.

References

[1] Ngoma C, Igira FT. Empowering community health workers to collect and record maternal and child health data by resolving contradictions. J Health Inform Afr. 2015;3(1):1-18.

[2] Ayumba EM. Modelling software agents: Web-based decision support system for malaria diagnosis and therapy. J Health Inform Afr. 2015;3(1):30-36.

[3] Yehualashet G, Andualem M, Tilahun B. The attitude towards and use of electronic medical record system by health professionals at a referral hospital in northern Ethiopia: Cross-sectional study. J Health Inform Afr. 2015;3(1):19-29.

[4] Croft WB, Metzler D, Strohman T. Search engines: information retrieval in practice. Harlow: Pearson Education; 2015.

[5] Manning CD, Raghavan P, Schütze H. An introduction to information retrieval. New York: Cambridge University Press; 2008.

[6] Goeuriot L, Kelly L, Jones GJF. Test collections for medical information retrieval evaluation. In: SIGIR 2013 Health Search and Discovery HSD workshop. Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval: 2013 Jul 28 - Aug 1; Dublin, Ireland. SIGIR; 2013.

[7] Gale EAM. Declassifying diabetes. Diabetologia. 2006;49:1989-1995.

[8] Gale EAM. The discovery of type 1 diabetes. Diabetes. 2001;50:217-226.

[9] Niklasson B, Samsioe A, Blixt M, Sandler S, Sjöholm A, Lagerquist E, et al. Prenatal viral exposure followed by adult stress produces glucose intolerance in a mouse model. Diabetologia. 2006;49:2192-2199.

[10] Alberti KGMM. Problems related to definitions and epidemiology of type 2 (non-insulin-dependent) diabetes mellitus: studies throughout the world. Diabetologia. 1993;36(10):978-984.

[11] Liao YH, Ko YL, Tsai LP. A rare genetic disorder in juvenile diabetes: Wolfram syndrome - case report. HKCPaed. 2015;20:172-175.

[12] Ascher H. Paediatric aspects of coeliac disease: old challenges and new ones. Digestive and Liver Disease. 2002;34:216-224.

[13] Majeed A, Newnham A, Ryan R, Khunti K. Prevention and cure of type 2 diabetes : General practitioners are treating more cases of diabetes. BMJ. 2002;325(7370):965-965.

[14] Graham R. Self-monitoring of blood glucose (SMBG): Considerations for intensive diabetes management. P&T. 2005;30(12).

[15] Douek IF, Gillespie KM, Dix RJ, Bingley PJ, Gale EAM. Three generations of autoimmune diabetes: an extended family study. Diabetologia. 2003;46:1313-1318.

[16] Jun HS, Yoon JW. A new look at viruses in type 1 diabetes. DMMR. 2002;19:8-31.

[17] Members of the Expert Committee. Report of the expert committee on the diagnosis and classification of diabetes mellitus. Diabetes Care. 2003;26(1):S5-S20.

[18] Dean L, McEntyre J. The genetic landscape of diabetes. Bethesda, MD: NCBI; 2004.

[19] American Diabetes Association. Standards of medical care in diabetes. Diabetes Care. 2010;33(1):S11-S61.

[20] Jones K, Saad R. Brief review of diabetes mellitus type 1. Morning Report Newsletter. 2012;1(6).

[21] Canadian Diabetes Association. Standards of care for students with type 1 diabetes in school. Standards of Care. 2008.

[22] National Diabetes Data Group. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. Diabetes. 1979;28:1039-1057.

[23] Kanungo A, Samal KC, Sanjeevi CB. Molecular mechanisms involved in the etiopathogenesis of malnutrition-modulated diabetes mellitus. In Sanjeevi B. Immunology of diabetes. New York: Academy of Sciences; 2002.

[24] Trudeau JD, Kelly-Smith C, Verchere B, Elliott JF, Dutz JP, Finegood DT, et al. Prediction of spontaneous autoimmune diabetes in NOD mice by quantification of autoreactive T cells in peripheral blood. JCI. 2003;111(2):217-223.

[25] Hill NJ, Stotland A, Solomon M, Secrest P, Getzoff E, Sarvetnick N. Resistance of the target islet tissue to autoimmune destruction contributes to genetic susceptibility in type 1 diabetes. Biology Direct. 2007;2(5):1-20.

[26] Van Bon AC, Kohinor MJE, Hoekstra JBL, Von Basum G, DeVries JH. Patients' perception and future acceptance of an artificial pancreas. JDST. 2010;4(3):596-602.

[27] Katahira M. A proposal for a new classification of Type 1 Diabetes Mellitus based on clinical and immunological evidence. Recent Pat Endocr Metab Immune Drug Discov. 2009;3:54-59.

[28] Hassan J. Overview on diabetes mellitus (type 2). Chromatography Separation Techniques. 2013;4(2).

[29] National Health Service. Diabetes, type 2. NHS [Internet] 2014 Jan. Available from: http://www.nhs.uk/conditions/diabetes-type2/pages/introduction.aspx

[30] Chougale AD, Panaskar SN, Gurao PM, Arvindekar AU. Optimization of alloxan dose is essential to induce stable diabetes for prolonged period. Asian Journal of Biochemistry. 2007;2(6):402-408.

[31] Gale EAM. The rise of childhood type 1 diabetes in the 20th century. Diabetes. 2002;51:3353-3361.

[32] Jali M, Kambar S. Prevalence of peripheral neuropathy in Indians with newly diagnosed type 2 diabetes. Proceedings of the 19th World Diabetes Congress; 2006 Dec 3-7; Cape Town, South Africa. IDF; 2006.

[33] Badran M, Laher I. Type II diabetes mellitus in Arabic-speaking countries. International Journal of Endocrinology. 2012;1-11.

[34] Kim A. Persistent RNA virus infection and development of type I diabetes [dissertation]. Calgary (AL): Calgary University; 2000.

[35] Landin-Olsson M. Latent autoimmune diabetes in adults. In Sanjeevi, B. Immunology of diabetes. New York: Academy of Sciences; 2002.

[36] National Diabetes Information Clearinghouse. The diabetes dictionary. Bethesda, MD: NDIC; 2009.

[37] Shtauvere-Brameus A. Studies of autoimmune diabetes in Latvians and other populations [dissertation]. Stockholm: Karolinska Institutet; 2001.

[38] Harris M. Classification, diagnostic criteria, and screening for Diabetes. In National Diabetes Information Clearinghouse. Diabetes in America. 2nd ed. Bethesda, MD: NDIC; 2013.

[39] Chatenoud L, Bach JF. Questioning four preconceived ideas on immunotherapy of clinical type 1 diabetes: lessons from recent CD3 antibody trials. RDS. 2005;2(3):116-120.

[40] Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Quarterly. 2004;28(1):75-105.

[41] Gregor S. The nature of theory in information systems. MIS Quarterly. 2006;30(3):611-642.

[42] Gregor S, Hevner AR. Positioning and presenting design science research for maximum impact. MIS Quarterly. 2013;27(2):337-355.

[43] Hevner AR. Keynote - Designing informing systems: What research tells us. Informing Science and IT Education Conferences; 2015 Jun 29-Jul 5; Tampa, FL, USA. InSITE; 2015.

[44] Harris ZS. The structure of science information. JBI. 2002;35:215-221.

[45] Singhal A. Modern information retrieval: A brief overview. IEEE Data Engineering Bulletin. 2001;24(4):35-43.

[46] Clarke CLA, Cormack GV, Tudhope EA. Relevance ranking for one to three term queries. IPM. 2000;36:291-311.