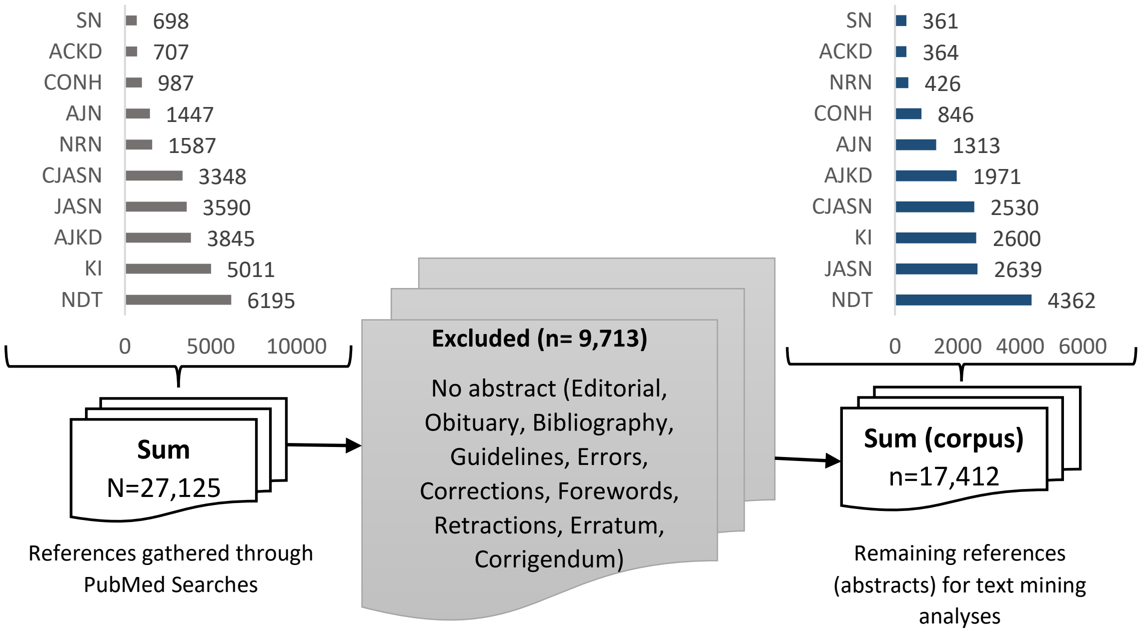
SUPPLEMENTAL

Methods

Phase 1: Establishment of the Corpus

In this phase, we identified and downloaded paper references from top ten nephrology journals (the ranking was based on SCimago Journal Rank indicators (SJI), a measure of journal's impact, and influence and prestige) into EndNote for the last ten years (2007-2017). As shown in Figure 1, the initial search generated 21,125 paper references. However, after removing the references that had no abstracts (such as editorials, book reviews, etc.), a total of 17,412 references with abstracts remained in the final corpus.



Supplemental Methods Figure 1: Flow diagram of included studies in text mining

*(**ACKD:** Advances in Chronic Kidney Disease, **AJKD:** American Journal of Kidney Disease, **AJN:** American Journal of Nephrology, **CJASN:** Clinical Journal of American Society of Nephrology, **CONH:** Current Opinions in Nephrology and Hypertension, **JASN:** Journal American Society of Nephrology, **KI:** Kidney International, **NDT:** Nephrology Dialysis and Transplant, **NRN:** National Review of Nephrology, **SN:** Seminars in Nephrology)*

A sample data record from the corpus, provided in Table 1, highlights some of the major fields that we obtained and utilized during the analyses. Even though the actual data included more fields such as paper title and keywords, we opted to analyze abstracts only due to the better predictive capability of the abstracts[1]. This is because abstracts better reflect the content of the paper than the title or keywords[2-4].

Supplemental Methods Table 1. Sample Data Record from the Corpus

Field Name	Sample Data
Year	2017
Author	Jardine, M. J.: Mahaffey, K. W.: Neal, B.: Agarwal, R.: Bakris, G. L.: Brenner, B. M.: Bull, S.: Cannon, C. P.: Charytan, D. M.: de Zeeuw, D.: Edwards, R.: Greene, T.: Heerspink, H. J. L.: Levin, A.: Pollock, C.: Wheeler, D. C.: Xie, J.: Zhang, H.: Zinman, B.: Desai, M.: Perkovic, V.[5]
Title	*A comparative evaluation of various methods for microalbuminuria screening*
Journal	The Canagliflozin and Renal Endpoints in Diabetes with Established Nephropathy Clinical Evaluation (CREDENCE) Study Rationale, Design, and Baseline Characteristics
Journal Abbr.	AJN
Abstract	BACKGROUND: People with diabetes and kidney disease have a high risk of cardiovascular events and progression of kidney disease. Sodium glucose co-transporter 2 inhibitors lower plasma glucose by reducing the uptake of filtered glucose in the kidney tubule, leading to increased urinary glucose excretion. They have been repeatedly shown to induce modest natriuresis and reduce HbA1c, blood pressure, weight, and albuminuria in patients with type 2 diabetes. However, the effects of these agents on kidney and cardiovascular events have not been extensively studied in patients with type 2 diabetes and established kidney disease. METHODS: The Canagliflozin and Renal Endpoints in Diabetes with Established Nephropathy Clinical Evaluation (CREDENCE) trial aims to compare the efficacy and safety of canagliflozin -versus placebo at preventing clinically important kidney and cardiovascular outcomes in patients with diabetes and established kidney disease. CREDENCE is a randomized, double-blind, event-driven, placebo-controlled trial set in in 34 countries with a projected duration of approximately 5.5 years and enrolling 4,401 adults with type 2 diabetes, estimated glomerular filtration rate >/=30 to <90 mL/min/1.73 m2, and albuminuria (urinary albumin:creatinine ratio >300 to </=5,000 mg/g). The study has 90% power to detect a 20% reduction in the risk of the primary outcome (alpha = 0.05), the composite of end-stage kidney disease, doubling of serum creatinine, and renal or cardiovascular death. CONCLUSION: CREDENCE will provide definitive evidence about the effects of canagliflozin on renal (and cardiovascular) outcomes in patients with type 2 diabetes and established kidney disease. TRIAL REGISTRATION: EudraCT number: 2013-004494-28; ClinicalTrials.gov identifier: NCT02065791.

Phase 2: Generation and Curation of the Term List.

The term list generation and curation phrase, the most time demanding of the three phases, involves iterative processes of application of NLP tools to the text to generate the most representative term list and thereby turn the text into a high-quality structured data for further analyses. Table 2 briefly describes the terminology of the NLP tools and how these tools were applied to the textual content iteratively during the term list generation and curation phase. The most time demanding task during this phase was to determine the phrases that would be added to the term list by using n-grams, n sequencing of adjacent words. To automate this part, a list of medical terminology -as a lookup table- that could include two or more words in each term was added. We curated this list by using a glossary of kidney disease terms from Cochrane Kidney and Transplant[6] and medical terminology list from MedicineNet.com[7] by filtering the terminology with two or more words. The remaining relevant and important phrases were manually incorporated into the term list.

As it is shown in Table 2, some of the terminologies needed recoding since they were referring to the same/similar concepts. However, we did not recode some of the very high frequency terms such as renal into kidney or serum creatinine into the estimated GFR or nephropathy into stages of kidney diseases. We wanted to examine whether researchers have been adopting the initiatives to streamline the kidney disease terminology such as the National Kidney Foundation's Kidney Disease Outcomes Quality Initiative (NKF KDOQI).[8]

Supplemental Methods Table 2. Terminology for the NLP tools used during term generation and curation process

Concept	Description	Application
Text Document	Any written document that can be analyzed with NLP tools	Example Text: Background: 500 fellows who have complimentary …
Token	The smallest unit of text (group of characters) corresponding to a concept, like a word in a given text	background, 500, fellows who have complimentary membership at american society of nephrology (asn) completed the survey the results indicated that the fellows seeks more research training that involves chronic kidney disease (ckd), antibody-mediated-rejection (abmr), and double-blind placebo-controlled trials.
Tokenization	Breaking text into tokens	
Regular expression (regex)	Language that describes patterns and rules in a text document. It can be useful to exclude/include certain patterns such as punctuation marks and numbers from the text-mining analyses	Adding 'dash' to the list of characters to match words with those embedded characters allows antibody-mediated-rejection to become a single token: [&'-]
Stop word	Tokens excluded from analysis	background, who, have, at, of, the, that
Recoding	Renaming tokens in order to group or ungroup them. Frequently used to indicate synonyms	Antihypertensive medications/drugs/agents → Antihypertensive medications Arteriovenous fistula/fistulae → Arteriovenous fistula
Stemming / Lemmatization	Reduction of tokens into their simplest form (i.e. roots of the words)	(educate, educated, educating, education, educational, educative, educator) → educ·
Phrase	Combination of a small number of tokens	american. societi. of nephrolog·
N-gram	N sequence of words that are frequently used adjacent to each other	N=6, meaning that up to six word that are frequently used adjacent to each other will be recognized by the NLP algorithm for researchers to make decision. e.g. national health and nutrition examination survey (NHANES)
Term	A token or a phrase	american. societi. of nephrolog·, ckd
Document	The unstructured text included in the analysis for a particular record	The abstract of a particular publication
Corpus	The collection of documents included in the analysis	All of the abstracts included in the analysis
Term Frequency-Inverse Document	A measure that shows the importance of a term in a document while considering the whole corpus. TF= (Number of	For example if the word hypertension mentioned 4 times in an abstract of 100 words then the term frequency would be 0.04 (4/100). Moreover, if the word hypertension mentioned in total of 100 documents in a corpus of 10,000 abstracts then

Concept	Description	Application
Frequency (TF-IDF)	Terms)/ (Total number of words in a document); IDF= Log[(Total number of documents in the whole corpus)/(Number of documents that a particular term mentioned in the whole corpus)]	the Inverse Document Frequency would be: log(10,000/100)= 2. Lastly, the TF-IDF would be: 0.04*2=0.08
Document Term Matrix (DTM)	The matrix where rows correspond to document, columns correspond to terms and each cell corresponds to values of analysis based on the weighing option	Frequencies or TF IDF values for each document/term pair

After finalizing the term list by using NLP tools provided in Table 2, a document by term matrix (DTM) was generated. This DTM had 17,412 documents as rows and 9,968 terms as columns, where the values of the cell were determined using *Term Frequency - Inverse Document Frequency* (TF-IDF) centered and scaled. Table 3 presents a DTM example where a sample of documents (i.e. abstracts) are shown as rows and terms are shown as columns.

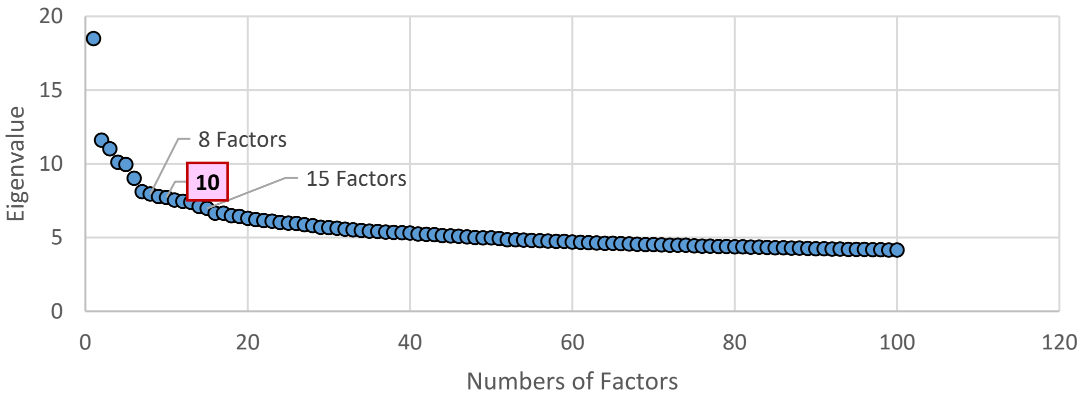
Supplemental Methods Table 3. Document by Term Matrix (DTM) Example

D	Abstract	ckd TF IDF	control· TF IDF	dialysis TF IDF	cell· TF IDF	protein· TF IDF
1	BACKGROUND: In rodent models…	0	0	0	1.813897	1.74367
2	TGF-beta1 expression…	0	0	0	0	0.871835
3	BACKGROUND/AIMS: Vascular…	0	0	0	0	0.871835
4	BACKGROUND: After insult to the kidney…	0	0	0	0	0
5	Because fibrotic kidneys exhibit aberrant...	0	0	0	0.906949	0
6	Enhancer of zeste homolog 2 (EZH2)…	1.578854	0	0	0	0.871835
7	BACKGROUND: Gentamicin, a widely used…	0	0	0	5.441692	2.615505
8	The hallmark of renal tubulointerstitial …	0	0	0	0.906949	0.871835
9	CKD is a major public health problem…	0.789427	0	0	0	0

D= Document, **TF** =Term Frequency, **IDF**= Inverse Document Frequency

Phase 3: Analyses of the Term List with Text Mining

In this phase, we first used Latent Semantic Analysis (LSA)[9-14], a dimension reduction method that resembles Principle Component Analysis. Specifically, we applied LSA to 17,412 x 9,968 DTM matrix where the cell values were determined using TF-IDF's centered and scaled. LSA utilizes Singular Value Decomposition (SVD) to reduce dimensions in DTM through series of linear approximations.[3] Through SVD, one can reduce the dimensions and identify the underlying major factors/topics in the data by capturing connections among terms. In order to find the optimum numbers of factors in the output of LSA, we generated a scree plot based on the principle components of the DTM matrix using 100 vectors (see Figure 2). As indicated by the eigenvalues in Figure 2, the first several factors accounted for a large proportion of the total variability in the data. Generally, the optimum numbers of factors would lie at the turning points in the scree plot. In our case, according to the scree plot, the optimum numbers of factors laid between 8 to 15 factors (Figure 2). In order to reveal the themes and determine the optimum number of factors, we then applied Topic Analyses (TA), which resembles the factor analyses with orthogonal Varimax rotation,[15] and generated topics from 8 to 15. We then manually determined 10 as the optimum numbers of topics (factors) for our data by examining 8 to 15 topics separately. As indicated by the eigenvalues in Figure 1, Factor 1 explains the highest proportion of the model followed by the other factors.



Supplemental Methods Figure 2: Finding optimum numbers of factors (topics) through scree plot that is based on a principle components of 9968 terms by 17412 documents using TF IDF weighting centered and scaled making 100 vectors.

As the final step in Phase 3, we generated the topics by identifying the best home for each of the 17,412 documents (abstracts) into the previously determined 10 topics. To achieve this, we identified the maximum topic score value that each document achieved under each of the 10 topics (see Table 4). Then, we assigned that particular document to the topic where it achieved the highest topic score. This process allowed us to generate 10 topics of documents that correspond to the 10 topics generated through TA.

Supplemental Methods Table 4. A Sample DTM for Topic Analysis and Assigning Each Abstract into Topics

D	A	T	Topic 1-4	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8-10	maxval	maxcol	T
1	…	…	…	5.335	-2.738	0.765	-1.023	…	5.335	Topic 4	4
2	…	…	…	5.032	7.877	0.068	-0.270	…	7.877	Topic 5	5
3	…	…	…	-0.517	-1.545	1.935	-2.274	…	1.935	Topic 6	6
4	…	…	…	1.230	-3.682	-3.846	8.304	…	8.304	Topic 7	7

A= Abstract**, T=** Topic**, D=** Document**, Maxcol=** the column (topic) that has the maximum numeric value in a row, **Maxval=** The maximum numeric value in each row

References

1. Chakraborty V, Chiu V, Vasarhelyi M. Automatic classification of accounting literature. *International Journal of Accounting Information Systems.* 2014;15(2):122-148.
2. Delen D, Crossland MD. Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications.* 2008;34(3):1707-1720.
3. Kim Y-M, Delen D. Medical informatics research trend analysis: A text mining approach. *Health Informatics Journal.* 2016;0(0):1460458216678443.
4. Miner G, Elder J, Fast A, Hill T, Delen D. *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications.* Academic Press; 2012.
5. Jardine MJM, K. W.: Neal, B.: Agarwal, R.: Bakris, G. L.: Brenner, B. M.: Bull, S.: Cannon, C. P.: Charytan, D. M.: de Zeeuw, D.: Edwards, R.: Greene, T.: Heerspink, H. J. L.: Levin, A.: Pollock, C.: Wheeler, D. C.: Xie, J.: Zhang, H.: Zinman, B.: Desai, M.: Perkovic, V. The Canagliflozin and Renal Endpoints in Diabetes with Established Nephropathy Clinical Evaluation (CREDENCE) Study Rationale, Design, and Baseline Characteristics. *American journal of nephrology.* 2017;46(6):462-472.
6. Cochrane. Cochrane Kidney and Transplant - Glossary of Kidney Disease Terms. https://kidneyandtransplant.cochrane.org/cochrane-renal-group-glossary. Published 2018. Accessed September 09, 2018.
7. MedicineNet. MedTerms Medical Dictionary A-Z List. https://www.medicinenet.com/script/main/alphaidx.asp?p=a_dict. Published 2018. Accessed September 06, 2018.
8. KDOQI N. NKF Kidney Disease Outcomes Quality Initiative (NKF KDOQI)™. https://www.kidney.org/professionals/guidelines. Published 2018. Accessed October 22, 2018.
9. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *Journal of the American society for information science.* 1990;41(6):391-407.
10. Evangelopoulos N, Zhang X, Prybutok VR. Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems.* 2012;21(1):70-86.
11. Hofmann T. Unsupervised Learning by Probabilistic Latent Semantic Analysis. *Machine Learning.* 2001;42(1):177-196.
12. Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review.* 1997;104(2):211.
13. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. *Discourse Processes.* 1998;25(2-3):259-284.
14. Landauer TK, Laham D, Derr M. From paragraph to graph: Latent semantic analysis for information visualization. *Proceedings of the National Academy of Sciences.* 2004;101(suppl 1):5214-5219.
15. Abdi H. Factor rotations in factor analyses. *Encyclopedia for Research Methods for the Social Sciences Sage: Thousand Oaks, CA.* 2003:792-795.