A Concept–Wide Association Study of Clinical Notes to Discover New Predictors of Kidney Failure

Karandeep Singh,[*†] Rebecca A. Betensky,[‡] Adam Wright,[§‖] Gary C. Curhan,[‖¶**] David W. Bates,[§‖††] and Sushrut S. Waikar[‖¶]

Abstract

Background and objectives Identifying predictors of kidney disease progression is critical toward the development of strategies to prevent kidney failure. Clinical notes provide a unique opportunity for big data approaches to identify novel risk factors for disease.

Design, setting, participants, & measurements We used natural language processing tools to extract concepts from the preceding year's clinical notes among patients newly referred to a tertiary care center's outpatient nephrology clinics and retrospectively evaluated these concepts as predictors for the subsequent development of ESRD using proportional subdistribution hazards (competing risk) regression. The primary outcome was time to ESRD, accounting for a competing risk of death. We identified predictors from univariate and multivariate (adjusting for Tangri linear predictor) models using a 5% threshold for false discovery rate (q value <0.05). We included all patients seen by an adult outpatient nephrologist between January 1, 2004 and June 18, 2014 and excluded patients seen only by transplant nephrology, with preexisting ESRD, with fewer than five clinical notes, with no follow-up, or with no baseline creatinine values.

Results Among the 4013 patients selected in the final study cohort, we identified 960 concepts in the unadjusted analysis and 885 concepts in the adjusted analysis. Novel predictors identified included high–dose ascorbic acid (adjusted hazard ratio, 5.48; 95% confidence interval, 2.80 to 10.70; $q<0.001$) and fast food (adjusted hazard ratio, 4.34; 95% confidence interval, 2.55 to 7.40; $q<0.001$).

Conclusions Novel predictors of human disease may be identified using an unbiased approach to analyze text from the electronic health record.

Clin J Am Soc Nephrol 11: 2150–2158, 2016. doi: 10.2215/CJN.02420316

*Division of Learning and Knowledge Systems, Department of Learning Health Sciences and [†]Division of Nephrology, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan; Departments of [‡]Biostatistics, **Epidemiology, and [††]Health Policy and Management, Harvard T.H. Chan School of Public Health, Boston, Massachusetts; [§]Division of General Internal Medicine, Department of Medicine and [¶]Division of Renal Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts; and [‖]Department of Medicine, Harvard Medical School, Boston, Massachusetts

Correspondence: Dr. Karandeep Singh, 1161H North Ingalls Building, 300 North Ingalls Street, Ann Arbor, MI 48109-5403. Email: kdpsingh@umich.edu

Introduction

Every year, over 100,000 Americans develop ESRD requiring RRT with hemodialysis, peritoneal dialysis, or kidney transplantation (1). Predictors of kidney disease have traditionally been identified through classic epidemiologic approaches, whereby individual risk factors are adjusted for known or suspected confounders and evaluated for their association with CKD progression. This has led to the identification of a number of potential associations with the development of ESRD (2–8), including race, comorbid conditions, conventional and novel biomarkers, lifestyle factors, and medications.

A limitation of retrospective or prospective cohort studies is that associations are typically tested with prespecified covariates. This is akin to candidate gene approaches used to study the genetic basis of diseases. Genome–wide association studies have enabled the discovery of new associations through a paradigm of simultaneous, unbiased testing of multiple associations. Similar approaches have been used to discover new associations between diseases and environmental exposures using environment–wide association studies (9,10) and between a single genetic variant and multiple phenotypes using phenome–wide associations studies (11).

Although work has been done using approaches, such as topic modeling, to incorporate unstructured notes into prediction models (12), modern epidemiologic approaches have not used the clinical narrative in the discovery of disease associations. Clinical notes may contain a rich description of numerous epidemiologic exposures. Discovering associations on the basis of the clinical narrative carries the added complexity of an open cohort, where patients may enter and leave the cohort at various time points and may be lost to follow-up or experience competing events.

In this study, we present and show a new methodology, which we term a concept–wide association study, for examining relationships between several thousand concepts extracted from clinical notes with the development of ESRD.

Materials and Methods

Study Design

We conducted a retrospective cohort study using the full text of clinical notes in the year before the first outpatient general nephrology visit. The date of the first outpatient general nephrology visit was identified

through visit and billing data. The outcome was defined using International Classification of Diseases, 9th revision diagnosis codes (585.6, V42.0, and 996.81) and procedure codes (90970, 50360, 50365, 50370, 50380, 55.61, and 55.69) as the date on which RRT was first performed. All identified events were adjudicated through chart review.

Population Studied

Patients were included if they were seen between January 1, 2004, and June 18, 2014 by an adult nephrologist at a Brigham and Women's Hospital–affiliated outpatient clinic. Exclusion criteria included visits only with transplant nephrology, known ESRD before the first nephrology visit, fewer than five clinical notes in the year preceding the first nephrologist visit, no documented follow-up, or no baseline creatinine values. The threshold of five notes was chosen, because 85% of patients had at least five notes in the year preceding the first nephrology visit.

Data Collection

We obtained data from the Partners Research Patient Data Registry, a centralized clinical data warehouse. We obtained information on patient demographics (age, sex, and race), billing codes, the full text of all electronic clinical notes, and laboratory values, including serum creatinine, calcium, phosphorus, albumin, bicarbonate, and the urine albumin-to-creatinine ratio. We defined baseline laboratory values for each laboratory test as the first available result on or after the first nephrology visit up to 365 days after the visit. Death was determined from the Social Security Death Index, and ascertainment was limited to 30 days after the final clinical note. If ESRD did not occur by this time, the observation was treated as censored. The Partners Healthcare Institutional Review Board approved this study, and the need for informed consent was waived. We adhered to the Declaration of Helsinki.

Extraction of Concepts from Clinical Notes

Concepts rather than individual words were evaluated so that phrases, such as "CHF" and "congestive heart failure" could be considered together when evaluating their association with kidney failure. Concepts were extracted from all clinical notes starting from 1 year before the first nephrologist visit up to but not including the nephrology visit (Figure 1). All clinical note types (including phone calls, outpatient notes, and inpatient notes) were included. Of note, Brigham and Women's Hospital's electronic health record is primarily an outpatient record; although it contains admission and initial consultation notes, it typically does not contain inpatient progress notes. Concepts were coded as binary variables for each patient. Concepts with <1% patient prevalence were not evaluated further.

Concepts were extracted from notes in a two-step process. First, negated phrases were removed using the NegEx negation engine (13) using Python 2.7 (14). Second, notes were processed with the National Library of Medicine's MetaMap software (15) (version 2013v2), which maps phrases to Unified Medical Language System codes known as concept unique identifiers. Extracted concepts were restricted to the Systematized Nomenclature of Medicine–Clinical Terms and RxNorm ontologies to map general and drug concepts, respectively. Concepts were not limited by semantic type, and therefore, all types of concepts contained were extracted (*e.g.*, diagnoses, medications, signs, and symptoms). Mapping of phrases to multiple concepts was allowed (*e.g.*, "heart failure" maps to concepts for both "heart failure" and "congestive heart failure").

Natural language processing systems may occasionally create erroneous mappings (*e.g.*, the phrase "GN" in a clinical note maps to the concept for "Guinea Republic"). Instead of reporting the name of the concept intended by the phrase mappings, we report the phrase (or phrases) matching with the highest frequency for each concept (*i.e.*, we report "GN" and not "Guinea Republic").

Statistical Analyses

Proportional subdistribution hazards (competing risk) regression as described by Fine and Gray (16) was performed with ESRD defined as the event of interest and death considered as a competing risk. The Fine and Gray (16) model directly assesses the effect of covariates on the cumulative incidence of a particular type of failure in a competing risks setting. The subdistribution hazard is the rate of ESRD per unit time for individuals who are still alive at that time or have died before that time. That is, individuals who have died without ESRD are treated as though they are still at risk for ESRD. Competing risk regression is preferable for prognostic questions, whereas Cox regression is preferable for etiologic questions (17). We were interested in identifying predictors that would differentiate patients whose kidney disease progressed to ESRD versus those who either did not progress or died—a question of prognosis—and therefore, we chose competing risk regression as our primary approach.

Competing risk regression for each concept was carried out in two phases: (*1*) unadjusted and (*2*) adjusted for a published ESRD risk prediction score developed by Tangri *et al.* (18) using age, race, sex, and baseline laboratory values (serum creatinine, calcium, phosphorus, albumin, bicarbonate, and the logarithm of the urine albumin-to-creatinine ratio).

Multiple hypothesis testing (19) was accounted for using the method by Storey (20), which controls the false discovery rate, defined as the expected proportion of false positives among all significant hypotheses. Using this method, P values were transformed into q values. Hazard ratios (HRs) and 95% confidence intervals (95% CIs) were not adjusted in any way. Concepts with q values <0.05 were reported as associations; this equates to a 5% expected proportion of false positives among all concepts declared to have associations. The false discovery rate method was chosen, because it explicitly controls the error rate of test conclusions among significant results, scales well in the face of increasing numbers of tests, and has higher power compared with the Bonferroni method (19). Concepts with HR>1 were identified as positive predictors, and concepts with HR<1 were identified as negative predictors. Analyses were performed in R 3.1.2 (21). q Values were computed using the qvalue R package (available on Bioconductor) by Storey *et al.* (22), and multiple imputation was performed using the mice R package (23).

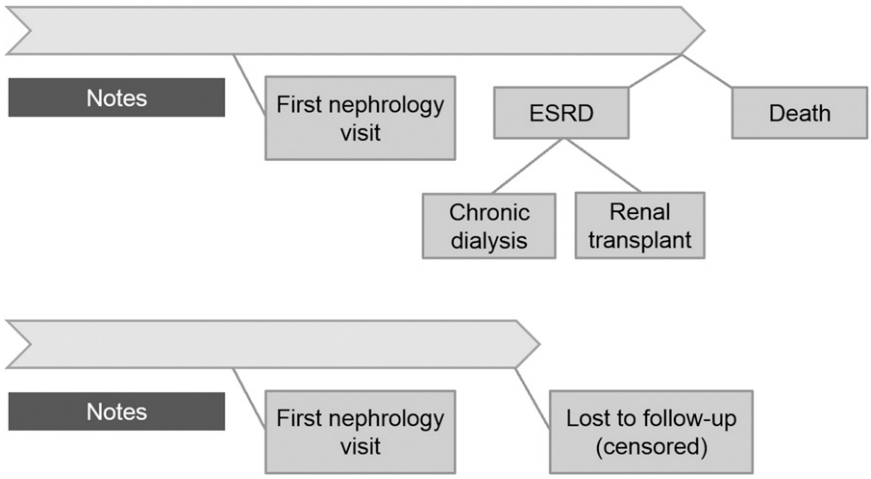

Figure 1. | Clinical notes from the year preceding the first nephrology visit were processed, and the extracted concepts were evaluated for associations with time-to-ESRD.

Identifying Associated Concepts

Each concept is likely to have a number of associations with other concepts in either the same direction (*e.g.*, diabetes mellitus and insulin) or opposite directions (*e.g.*, women and men). For a given concept, knowing its associated concepts may help in evaluating its plausibility as a true risk factor or a confounder. The Φ-coefficient is a measure of association for two binary variables, and it can be derived by computing a Pearson correlation on binary variables. We reported the three phrases with the greatest Φ-coefficients (in either direction) for each concept identified in our analysis after excluding duplicates.

Handling Missing Data

The adjusted analysis required calculation of the Tangri score, which is dependent on several variables (18). Missing values were multiply imputed using regression switching with predictive mean matching, a nonparametric method (24). Five datasets were imputed incorporating the event flag, log time, age, race, sex, serum creatinine, calcium, phosphorus, albumin, bicarbonate, and log urine albumin-to-creatinine ratio. The Tangri score was calculated using the imputed variables. Results were pooled using the rules by Rubin (25).

Sensitivity Analyses

The analysis was repeated using Cox regression with death considered as a censoring event. Death-censored ESRD has been used by several ESRD prediction models but may be problematic for the purpose of estimating the risk of ESRD, because death likely represents informative censoring (8,18,26–30). The hazards from a cause–specific hazard model are conditional on survival and cannot be interpreted as marginal hazards that ignore death. If death is independent of ESRD, results similar to those obtained by competing risk regression would be expected with the cause–specific Cox regression. However, in settings where a variable of interest is associated with a competing event, the two approaches may yield conflicting results (31). The adjusted competing risk and Cox analyses were repeated using nonmissing covariates (age, sex, black race, and eGFR) to limit bias from imputation.

Results

After applying inclusion and exclusion criteria to 9817 patients seen in nephrology clinic, we identified 4013 patients who constituted the analytic cohort (Figure 2). Of these, 134 were confirmed to have developed ESRD during follow-up, and 160 were confirmed to have died without developing ESRD (Table 1). Median follow-up time was 1.2 years (range =0–10.2 years).

After processing 103,962 clinical notes authored by 4589 distinct clinicians with NegEx and MetaMap, 38,698 unique concepts were identified. Of these, 7576 were present in at least 1% of patients and subsequently studied.

Predictors of ESRD from Univariate Analysis

Using a false discovery rate threshold of <5% ($q<0.05$) (Table 2), competing risk regression identified 960 concepts with q values <0.05 (Supplemental Table 1). Of these, 184 had an HR>1, and 776 had an HR<1.

Of the positive predictors included in the Tangri score (18)—a CKD risk prediction score on the basis of demographic and laboratory data—the unadjusted analysis identified men ("man"), "creatinine," and "proteinuria" as concepts associated with higher risk of ESRD. Mention of "proteinuria" had an HR of 2.16 (95% CI, 1.51 to 3.10; $q<0.001$), and "nephrotic-range proteinuria" had an HR of 5.09 (95% CI, 2.48 to 10.40; $q<0.001$). Other notable positive predictors included concepts related to heart failure; coronary artery disease; type 1 diabetes; complications of chronic kidney, such as hyperkalemia and anemia; medications associated with each of these conditions; and noncompliance with medications. Interestingly, the mention of "no swelling" was also associated with ESRD.

Negative predictors for ESRD included a large number of factors that may be indicative of either a positive association with death (*e.g.*, "ventilator"), poor candidacy for dialysis, or kidney transplant (*e.g.*, "metastatic cancer")

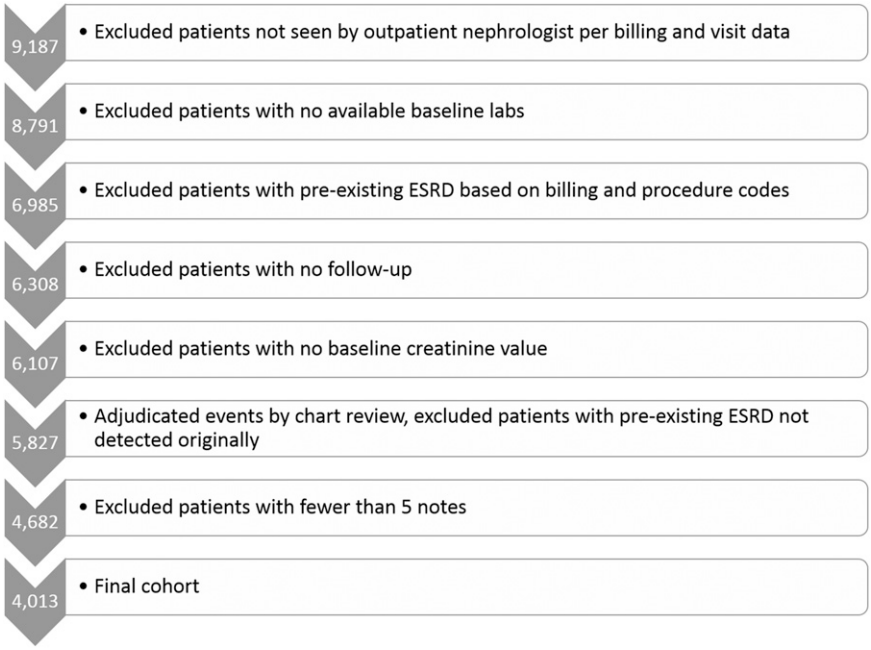

Figure 2. | Approximately half of the patients considered were included in the study cohort.

or a lower risk of ESRD (*e.g.*, "female"). Although we cannot distinguish among these reasons for a given concept, all contribute negatively to the cumulative incidence of progression to ESRD in the framework of a competing risk model. Interestingly, "female" had an HR of 0.58 (95% CI, 0.41 to 0.83; q=0.02), which is a near-perfect reciprocal of the risk conferred by "male," whereas "healthful food" had an HR of <0.01 (95% CI, <0.01 to <0.01;

Table 1. Baseline characteristics of the patients included in the study and patients excluded on the basis of having fewer than five clinical notes in the year before the first nephrology visit

Characteristic	No. (%) or Mean (SD)	
	All Patients Included in Study	Patients Excluded for Having Fewer Than Five Clinical Notes in 1 yr before First Nephrology Visit
No. of patients	4013 (100%)	669 (100%)
Age, yr	59.3 (16.6)	53.9 (16.8)
Women	2347 (58.4%)	348 (52.0%)
Black	568 (14.2%)	103 (15.4%)
Baseline eGFR calculated by CKD-EPI formula, ml/min per 1.73 m^2		
≥90	1014 (25.3%)	171 (25.6%)
60–89	1225 (30.5%)	162 (24.2%)
30–59	1263 (31.5%)	180 (26.9%)
15–29	363 (9.0%)	94 (14.1%)
<15	148 (3.7%)	62 (9.3%)
Microalbumin-to-creatinine ratio, μg/mg, median (range)	33.4 (<30–12,510)	75.3 (<30–7932)
Time before first nephrology visit, yr, median (range)	5.4 (0–10.2)	1.0 (0–10.0)
Follow-up time after nephrology visit, yr, median (range)	1.2 (0–10.2)	2.3 (0–10.3)
No. of notes in the year before first nephrology visit, median (range)	19 (5–293)	3 (1–4)
No. of concepts extracted per patient, median (range)	745 (3–4062)	0 (0–515)

CKD-EPI, Chronic Kidney Disease Epidemiology Collaboration.

Table 2. Unadjusted competing risk regression for ESRD showing concepts with q values <0.05 (top 10 positive and negative predictors shown; full results are in Supplemental Table 1)

UMLS CUI	Most Common Phrase (Three Most Correlated Phrases [Φ-Coefficient])	HR [95% CI]	q Value	Prevalence (%)
Top 10 positive predictors (HR>1) by q value (lowest to highest)				
C0022661	Chronic renal insufficiency (CKD [0.77], chronic disease [0.71], renal insufficiency [0.66])	5.07 [3.52 to 7.31]	<0.001	15.9
C0022658	Nephropathy (renal impairment [0.89], renal insufficiency [0.81], CKD [0.7])	3.45 [2.44 to 4.89]	<0.001	17.2
C0917873	Dialysis[a] (hemodialysis [0.38], ESRD [0.34], uremia [0.3])	4.64 [3.01 to 7.17]	<0.001	5.6
C0011946	Dialysis[a] (hemodialysis [0.37], ESRD [0.34], uremia [0.3])	4.64 [3.01 to 7.16]	<0.001	5.6
C0403447	CKD (chronic disease [0.92], renal impairment [0.78], chronic renal insufficiency [0.77])	3.36 [2.38 to 4.73]	<0.001	21.9
C0035078	Renal insufficiency (renal impairment [0.91], nephropathy [0.81], chronic renal insufficiency [0.66])	3.55 [2.47 to 5.11]	<0.001	25.4
C0546817	Volume overload (diuresis [0.61], diuresed [0.6], BNP [0.52])	3.85 [2.54 to 5.84]	<0.001	7.2
C0020223	Hydralazine (hydralazine HCl [0.65], labetalol [0.33], CK-MB [0.29])	4.01 [2.54 to 6.33]	<0.001	4.4
C0699992	Lasix (furosemide [0.74], sulfonamide [0.61], furosemide 40 mg [0.51])	2.82 [2.00 to 3.97]	<0.001	19
C0016860	Furosemide (sulfonamide [0.8], lasix [0.74], furosemide 40 mg [0.66])	2.84 [1.99 to 4.07]	<0.001	13.6
Top 10 negative predictors (HR<1) by q value (lowest to highest)				
C0227391	Sigmoid (descending colon [0.43], transverse colon [0.36], large bowel [0.33])	<0.01 [<0.01 to <0.01]	<0.001	6.6
C1299496	Radiology report (anion gap [0.14], 3.9 [0.14], pigment [0.13])	<0.01 [<0.01 to <0.01]	<0.001	4.1
C0231953	Lung volumes (FRC [0.46], diffusing capacity [0.45], atelectasis [0.36])	<0.01 [<0.01 to <0.01]	<0.001	4.6
C0014672	Equipment (procedure/site [0.42], knowledge deficit [0.4], nursing care [0.36])	<0.01 [<0.01 to <0.01]	<0.001	4.7
C0221103	Suppression (acid [0.24], suppressed [0.17], bone marrow [0.15])	<0.01 [<0.01 to <0.01]	<0.001	4.1
C0439181	Suppression (acid [0.24], suppressed [0.17], bone marrow [0.15])	<0.01 [<0.01 to <0.01]	<0.001	4.1
C0086930	Risk assessment (homicidal ideation [0.73], abnormal movements [0.65], suicidal ideation [0.65])	<0.01 [<0.01 to <0.01]	<0.001	4.5
C0184924	Punch biopsy (punch biopsy [0.99], punch [0.58], suture removal [0.57])	<0.01 [<0.01 to <0.01]	<0.001	3.4
C0700201	Sleep disturbance (fear of falling [0.55], skin problems [0.51], sleep disorder [0.43])	<0.01 [<0.01 to <0.01]	<0.001	4.4
C0014180	Endometrium (myometrium [0.63], right ovary [0.5], right ovary [0.5])	<0.01 [<0.01 to <0.01]	<0.001	3.3

UMLS CUI, Unified Medical Language System concept unique identifier; HR, hazard ratio; 95% CI, 95% confidence interval; BNP, B-type Natriuretic Peptide; CK-MB, creatine kinase-myocardial band; FRC, functional residual capacity.

[a]Note that the phrase dialysis maps to two concepts, because we used settings in MetaMap that allowed a phrase to match multiple concepts (instead of choosing only one).

$q<0.001$), which is in the opposite direction of the risk conferred by "fast food" (HR, 3.62; 95% CI, 1.72 to 7.63; $q=0.005$) but with a much larger effect size.

Predictors of ESRD after Adjusting for Tangri Score

After adjusting for the Tangri score, we identified 885 concepts using a false discovery rate threshold of <5% ($q<0.05$) (Table 3), of which 130 had HRs>1 and 755 had HRs<1 using competing risk regression (Supplemental Table 3). Seven hundred seventy-two concepts were identified in both the univariate and multivariate competing risk analyses.

Missing Data

Among variables required for imputation of the Tangri score, the proportion of missingness in the final cohort was 56.1% for the urine albumin-to-creatinine ratio, 37.8% for phosphorus, 9.8% for serum albumin, 2.6% for calcium, 2.4% for bicarbonate, and 0% for age, sex, and eGFR.

Sensitivity Analyses

Using Cox regression models, 129 concepts were found to have $q<0.05$ in the univariate analysis: 120 concepts with HRs>1 and nine concepts with HRs<1 (Supplemental Table 2); 127 concepts were identified by both competing risk and Cox regression, 833 were by competing risk regression only, and 2 were by Cox regression only. After adjusting for the Tangri score using Cox regression, no concepts were found to have $q<0.05$.

The competing risk–adjusted analysis was repeated using nonmissing covariates. Adjusting for age, sex, race, and eGFR, 843 concepts were identified with $q<0.05$, 95 were identified with HRs>1, and 748 were identified with HRs<1 (Supplemental Table 4). Of these, 791 concepts were identified in both the primary analysis and this analysis, 52 concepts were identified in this analysis only, and 94 concepts were identified in the primary analysis only. Cox regression with this set of covariates identified seven concepts (Supplemental Table 5), all with HRs>1, in contrast to no concepts identified in the primary analysis.

Discussion

This is the first study to use an unbiased approach using text from the clinical notes to identify predictors of human disease. The approach's face validity was confirmed by the identification of several well established risk factors for ESRD, including men, degree of proteinuria, diabetic kidney disease, heart failure, and anemia. This study also identified novel predictors of ESRD that have not been previously described, such as fast food and high–dose ascorbic acid.

This paper describes a hypothesis-generating approach, and caution is needed when interpreting the results. Identified predictors could fall into any of seven possible categories.

1. Vague and unlikely to be meaningful
2. Confounded by indication
3. Mention of the concept in a note, rather than its actual presence, indicates a risk (a specific type of confounding by indication)
4. False positive result (due to 5% false discovery rate)
5. Associated with the event of interest but not causal
6. Associated with the competing event
7. True risk factor (associated with the event of interest and causal)

Confounding by Indication

Confounding may be obvious in some cases (docusate as a marker of hospitalization) and less obvious in others. The association of pes planus (flat foot) or volar (sole of the foot) with ESRD may be confounded by documentation of a foot examination among diabetic individuals referred to podiatry for diabetic nephropathy. Of the 54 patients noted to have a finding of flat foot, 39 had diabetes mellitus mentioned their notes. The association of "Ascorbic Acid 500 MG" with ESRD may be confounded by its association with "cardiac transplantation"—a relationship that was identified through a systematic evaluation of interconcept associations—although the Φ-coefficient of 0.28 reflects only a weak association between the two concepts. These situations are akin to the linkage disequilibrium problem observed in genome–wide association studies, where false associations may be identified when two genes are in close proximity.

Biologically Plausible Predictors

High–dose ascorbic acid could conceivably lead to a higher risk of ESRD through its metabolism to oxalate, the most common constituent of kidney stones (a risk factor for CKD progression [32,33]) and the metabolic abnormality found in primary hyperoxaluria, a group of rare genetic diseases associated with kidney failure. Fast food is known to be high in sodium and phosphate content (34), and both excessive salt intake and hyperphosphatemia are known to blunt the renoprotective effects of angiotensin–converting enzyme inhibitors and promote CKD progression (35,36). However, neither of these concepts were identified in the adjusted Cox regression analysis, and therefore, the higher cumulative incidence of ESRD for these two concepts may in part be driven by a lower risk of death. In both instances, the lower risk of death may be driven by confounding due to healthy user effect (37).

Sensitivity to Statistical Approach

Cox regression identified fewer predictors than competing risk regression in both the unadjusted and adjusted analyses. The likely explanation for the observed differences is that Cox regression and competing risk regression ask two different but related questions. Cox regression tests whether the mention of a concept is directly associated with the outcome of ESRD, and competing risk regression tests whether individuals with a given concept present in their notes were more likely to (live to) experience ESRD (31). If a concept is associated with a higher risk of death (the competing event), the HR for ESRD as measured by competing risk regression will be lower than the cause-specific hazard, because fewer individuals are alive and able to experience ESRD and *vice versa*. Because many of the concepts have varying associations with death, it is not surprising that concepts that have either a positive or negative association with death are identified with competing risk regression but not with Cox regression.

Table 3. Competing risk regression for ESRD adjusted for Tangri score showing concepts with *q* values <0.05 (top 10 positive and negative predictors shown; full results are in Supplemental Table 3)

UMLS CUI	Most Common Phrase (Three Most Correlated Phrases [Φ-Coefficient])	HR [95% CI]	*q* Value	Prevalence (%)
Top 10 positive predictors (HR>1) by *q* value (lowest to highest)				
C0344355	Fast food (wheat [0.18], chicken [0.17], chips [0.17])	4.34 [2.55 to 7.40]	<0.001	1.4
C0455567	History of psoriasis (psoriasis [0.41], clobetasol [0.18], betamethasone [0.18])	6.00 [3.11 to 11.60]	<0.001	1.1
C1531390	Transfer of care (hyperactivity [0.14], transferred [0.14], imipenem [0.13])	5.15 [2.78 to 9.57]	<0.001	1.3
C0205133	Redacted name[a] (acute pain [0.13], callus [0.12], squats, squatting [0.12])	4.35 [2.46 to 7.70]	<0.001	1.3
C0983829	Ascorbic acid 500 mg (vitamin C [0.35], cardiac transplantation [0.28], tocopherol-DL-α [0.27])	5.48 [2.80 to 10.70]	<0.001	1.2
C0030704	Transferred care, transferring care (transferred [0.19], fentanyl citrate [0.16], sodium chloride 0.9 [0.16])	4.41 [2.43 to 8.02]	<0.001	1.7
C0412694	Liver MRI (α-fetoprotein [0.42], portal hypertension [0.39], HCC [0.36])	7.96 [3.37 to 18.80]	<0.001	1.8
C0991537	Oral susp (suspension [0.59], susp [0.45], atovaquone [0.44])	3.96 [2.20 to 7.14]	<0.001	2.4
C0443349	Volar (splint [0.24], wrist [0.2], wrist pain [0.19])	3.94 [2.15 to 7.20]	<0.001	1.4
C0023903	Liver tumor (α-fetoprotein [0.4], HCC [0.4], liver ultrasound [0.38])	7.49 [3.08 to 18.20]	<0.001	1.3
Top 10 negative predictors (HR<1) by *q* value (lowest to highest)				
C0536495	Moxifloxacin (oxacillin [0.25], tigecycline [0.24], linezolid [0.22])	<0.01 [<0.01 to <0.01]	<0.001	1.5
C0232721	Large stool (formed [0.17], normal stools [0.16], Dulcolax [0.15])	<0.01 [<0.01 to <0.01]	<0.001	1.1
C0429028	QT interval (PR interval [0.27], QRS duration [0.27], QRS axis [0.19])	<0.01 [<0.01 to <0.01]	<0.001	1.3
C0587461	Department dermatology (department hospital, hospital department [0.63], skin examination [0.19], dermatology [0.19])	<0.01 [<0.01 to <0.01]	<0.001	1
C1305830	CO2 15 (acidosis [0.25], pulmonary edema [0.2], ICU [0.2])	<0.01 [<0.01 to <0.01]	<0.001	1.4
C0337226	Fall home, fell home, fell in home (risk fall, risk falls [0.2], elder [0.19], unsteady [0.19])	<0.01 [<0.01 to <0.01]	<0.001	1.2
C2722359	s 2 (LV mass [0.51], ACC [0.39], pericardium/pleura [0.37])	<0.01 [<0.01 to <0.01]	<0.001	1.4
C0023911	Liver transplant (esophageal varices [0.4], portal hypertension [0.39], cirrhosis [0.36])	<0.01 [<0.01 to <0.01]	<0.001	1.3
C0040746	Liver transplant (esophageal varices [0.41], portal hypertension [0.39], cirrhosis [0.36])	<0.01 [<0.01 to <0.01]	<0.001	1.2
C0429012	QRS axis (12-lead ECG, 12-lead EKG [0.45], QRS duration [0.37], PR interval [0.37])	<0.01 [<0.01 to <0.01]	<0.001	1.4

UMLS CUI, Unified Medical Language System concept unique identifier; HR, hazard ratio; 95% CI, 95% confidence interval; MRI, magnetic resonance imaging; HCC, hepatocellular carcinoma; susp, suspension; CO2 15, carbon dioxide level of 15 mEq/L; ICU, intensive care unit; s 2, second heart sound; LV, left ventricular; ACC, accession number; ECG, electrocardiogram; EKG, electrocardiogram.
[a]Note that the redacted name was a reference to a podiatrist.

To a lesser extent, differences in results were also observed when choosing covariates that did not require imputation. Because over one half of patients had missing values for albuminuria and albuminuria is likely to be monitored more closely in patients with severe or worsening kidney disease, there are likely to be differences in patients with and without missing values. Multiple imputation is fairly robust to this problem as long as the plausible contributors to missingness are included in the imputation process, but we cannot know this definitively.

Limitations

The primary limitation of this study is that its findings are drawn from a single tertiary care center, which may have idiosyncrasies in documentation style and patient characteristics that may differ from other institutions. Validating this analysis in other cohorts is needed. Other limitations include missing covariate information and the open cohort design.

This approach was successful in translating the clinical narrative into a tool for the discovery of possible predictors that have not been previously linked to kidney failure. Future studies to replicate our findings and approach would be informative. The approach outlined here could potentially be used for patient-level prognosis, for population health management, or as a tool to identify previously unsuspected risk factors for CKD progression (38). As the adoption of electronic health records continues to rise and a generation of individuals has their entire health histories stored electronically, this approach provides a novel way to gain potential insights about disease risk as a natural byproduct of care delivery and electronic health record documentation.

Acknowledgments

Because Dr. Curhan is the Editor-in-Chief of CJASN, he was not involved in the peer-review process for this manuscript. Another editor oversaw the peer-review and decision-making process for this manuscript.

This research was supported, in part, by a National Institutes of Health T32 training grant awarded to the Division of Renal Medicine at Brigham and Women's Hospital.

The funding source had no role in the study design, conduct, analysis, or decision to submit the manuscript.

Disclosures

All authors have completed and submitted the International Committee of Medical Journal Editors Form for Disclosure of Potential Conflicts of Interest. D.W.B. is a coinventor on Patent No. 6029138 held by Brigham and Women's Hospital on the use of decision support software for medical management licensed to the Medicalis Corporation (Toronto, ON, Canada). He holds a minority equity position in the privately held company Medicalis Corporation, which develops web–based decision support for radiology test ordering. He serves on the board for SEA Medical Systems (San Jose, CA), which makes intravenous pump technology. He is on the clinical advisory board for Zynx, Inc. (Los Angeles, CA), which develops evidence-based algorithms. He consults for EarlySense (Ramat Gan, Israel), which makes patient safety monitoring systems. He receives equity and cash compensation from QPID, Inc. (Boston, MA), a company focused on intelligence systems for electronic health records. He receives cash compensation from CDI (Negev), Ltd. (Beersheba, Israel), which is a not for profit incubator for health information technology startups. He receives equity from Enelgy (Northridge, CA), which makes software to support evidence–based clinical decisions. He receives equity from Ethosmart (Ein Iron, Israel), which makes mobile applications to help patients with chronic diseases. He receives equity from Intensix (Netanya, Israel), which makes software to support clinical decision making in intensive care. He receives equity from MDClone (Beersheba, Israel), which takes clinical data and produces deidentified versions of it. The financial interests of D.W.B. have been reviewed by Brigham and Women's Hospital and Partners HealthCare in accordance with their institutional policies. Otherwise, no conflicts of interest were reported.

References

1. United States Renal Data System: *2014 USRDS Annual Data Report: An Overview of the Epidemiology of Kidney Disease in the United States*, Bethesda, MD, United States Renal Data System, 2014
2. Ejerblad E, Fored CM, Lindblad P, Fryzek J, Dickman PW, Elinder C-G, McLaughlin JK, Nyrén O: Association between smoking and chronic renal failure in a nationwide population-based case-control study. *J Am Soc Nephrol* 15: 2178–2185, 2004
3. Orth SR, Hallan SI: Smoking: A risk factor for progression of chronic kidney disease and for cardiovascular morbidity and mortality in renal patients–absence of evidence or evidence of absence? *Clin J Am Soc Nephrol* 3: 226–236, 2008
4. Perry HM Jr., Miller JP, Fornoff JR, Baty JD, Sambhi MP, Rutan G, Moskowitz DW, Carmody SE: Early predictors of 15-year end-stage renal disease in hypertensive patients. *Hypertension* 25: 587–594, 1995
5. Echouffo-Tcheugui JB, Kengne AP: Risk models to predict chronic kidney disease and its progression: A systematic review. *PLoS Med* 9: e1001344, 2012
6. Tangri N, Kitsios GD, Inker LA, Griffith J, Naimark DM, Walker S, Rigatto C, Uhlig K, Kent DM, Levey AS: Risk prediction models for patients with chronic kidney disease: A systematic review. *Ann Intern Med* 158: 596–603, 2013
7. Perneger TV, Whelton PK, Klag MJ: Risk of kidney failure associated with the use of acetaminophen, aspirin, and nonsteroidal antiinflammatory drugs. *N Engl J Med* 331: 1675–1679, 1994
8. Desai AS, Toto R, Jarolim P, Uno H, Eckardt K-U, Kewalramani R, Levey AS, Lewis EF, McMurray JJV, Parving H-H, Solomon SD, Pfeffer MA: Association between cardiac biomarkers and the development of ESRD in patients with type 2 diabetes mellitus, anemia, and CKD. *Am J Kidney Dis* 58: 717–728, 2011
9. Patel CJ, Bhattacharya J, Butte AJ: An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* 5: e10746, 2010
10. Patel CJ, Ioannidis JPA: Studying the elusive environment in large scale. *JAMA* 311: 2173–2174, 2014
11. Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, Field JR, Pulley JM, Ramirez AH, Bowton E, Basford MA, Carrell DS, Peissig PL, Kho AN, Pacheco JA, Rasmussen LV, Crosslin DR, Crane PK, Pathak J, Bielinski SJ, Pendergrass SA, Xu H, Hindorff LA, Li R, Manolio TA, Chute CG, Chisholm RL, Larson EB, Jarvik GP, Brilliant MH, McCarty CA, Kullo IJ, Haines JL, Crawford DC, Masys DR, Roden DM: Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 31: 1102–1110, 2013
12. Lehman LW, Saeed M, Long W, Lee J, Mark R: Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annu Symp Proc* 2012: 505–511, 2012
13. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG: A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 34: 301–310, 2001
14. Python Software Foundation: *Python Language Reference, version 2.7*, Wilmington, DE, Python Software Foundation, 2013
15. Aronson AR: Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. *Proc AMIA Symp* 2001: 17–21, 2001
16. Fine JP, Gray RJ: A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc* 94: 496–509, 1999

17. Noordzij M, Leffondré K, van Stralen KJ, Zoccali C, Dekker FW, Jager KJ: When do we need competing risks methods for survival analysis in nephrology? *Nephrol Dial Transplant* 28: 2670–2677, 2013
18. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, Levin A, Levey AS: A predictive model for progression of chronic kidney disease to kidney failure. *JAMA* 305: 1553–1559, 2011
19. Glickman ME, Rao SR, Schultz MR: False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *J Clin Epidemiol* 67: 850–857, 2014
20. Storey JD: The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann Stat* 31: 2013–2035, 2003
21. R Core Team: *R: A Language and Environment for Statistical Computing*, Vienna, Austria, R Core Team, 2014
22. Storey JD, Bass AJ, Dabney A, Robinson D: qvalue: Q-Value Estimation for False Discovery Rate Control, 2015. Available at: http://qvalue.princeton.edu. Accessed March 1, 2016
23. van Buuren S, Groothuis-Oudshoorn K: {mice}: Multivariate Imputation by Chained Equations in R. *J Stat Softw* 45: 1–67, 2011
24. Marshall A, Altman DG, Holder RL: Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: A resampling study. *BMC Med Res Methodol* 10: 112, 2010
25. Rubin DB: *Multiple Imputation for Nonresponse in Surveys*, New York, Wiley, 1987
26. Keane WF, Zhang Z, Lyle PA, Cooper ME, de Zeeuw D, Grunfeld J-P, Lash JP, McGill JB, Mitch WE, Remuzzi G, Shahinfar S, Snapinn SM, Toto R, Brenner BM; RENAAL Study Investigators: Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: The RENAAL study. *Clin J Am Soc Nephrol* 1: 761–767, 2006
27. Goto M, Wakai K, Kawamura T, Ando M, Endoh M, Tomino Y: A scoring system to predict renal outcome in IgA nephropathy: A nationwide 10-year prospective cohort study. *Nephrol Dial Transplant* 24: 3068–3074, 2009
28. Wakai K, Kawamura T, Endoh M, Kojima M, Tomino Y, Tamakoshi A, Ohno Y, Inaba Y, Sakai H: A scoring system to predict renal outcome in IgA nephropathy: From a nationwide prospective study. *Nephrol Dial Transplant* 21: 2800–2808, 2006
29. Johnson ES, Thorp ML, Platt RW, Smith DH: Predicting the risk of dialysis and transplant among patients with CKD: A retrospective cohort study. *Am J Kidney Dis* 52: 653–660, 2008
30. Landray MJ, Emberson JR, Blackwell L, Dasgupta T, Zakeri R, Morgan MD, Ferro CJ, Vickery S, Ayrton P, Nair D, Dalton RN, Lamb EJ, Baigent C, Townend JN, Wheeler DC: Prediction of ESRD and death among people with CKD: The Chronic Renal Impairment in Birmingham (CRIB) prospective cohort study. *Am J Kidney Dis* 56: 1082–1094, 2010
31. Lau B, Cole SR, Gange SJ: Competing risk regression models for epidemiologic data. *Am J Epidemiol* 170: 244–256, 2009
32. Baxmann AC, De O G Mendonça C, Heilberg IP: Effect of vitamin C supplements on urinary oxalate and pH in calcium stone-forming patients. *Kidney Int* 63: 1066–1071, 2003
33. Alexander RT, Hemmelgarn BR, Wiebe N, Bello A, Morgan C, Samuel S, Klarenbach SW, Curhan GC, Tonelli M; Alberta Kidney Disease Network: Kidney stones and kidney function loss: A cohort study. *BMJ* 345: e5287, 2012
34. Ritz E, Hahn K, Ketteler M, Kuhlmann MK, Mann J: Phosphate additives in food–a health risk. *Dtsch Arztebl Int* 109: 49–55, 2012
35. D'Elia L, Rossi G, Schiano di Cola M, Savino I, Galletti F, Strazzullo P: Meta-analysis of the effect of dietary sodium restriction with or without concomitant renin-angiotensin-aldosterone system-inhibiting treatment on albuminuria. *Clin J Am Soc Nephrol* 10: 1542–1552, 2015
36. Zoccali C, Ruggenenti P, Perna A, Leonardis D, Tripepi R, Tripepi G, Mallamaci F, Remuzzi G; REIN Study Group: Phosphate may promote CKD progression and attenuate renoprotective effect of ACE inhibition. *J Am Soc Nephrol* 22: 1923–1930, 2011
37. Shrank WH, Patrick AR, Brookhart MA: Healthy user and related biases in observational studies of preventive interventions: A primer for physicians. *J Gen Intern Med* 26: 546–550, 2011
38. Lazarus B, Chen Y, Wilson FP, Sang Y, Chang AR, Coresh J, Grams ME: Proton pump inhibitor use and the risk of chronic kidney disease. *JAMA Intern Med* 176: 238–246, 2016

Received: March 4, 2016 **Accepted:** September 1, 2016

Published online ahead of print. Publication date available at www.cjasn.org.

This article contains supplemental material online at http://cjasn.asnjournals.org/lookup/suppl/doi:10.2215/CJN.02420316/-/DCSupplemental.