Published in final edited form as:

Pharmacoepidemiol Drug Saf. 2011 November ; 20(11): 1138–1149. doi:10.1002/pds.2203.

Validation of The Health Improvement Network (THIN) Database for Epidemiologic Studies of Chronic Kidney Disease

Michelle R. Denburg, MD¹, Kevin Haynes, PharmD, MSCE², Justine Shults, PhD², James D. Lewis, MD, MSCE^{2,3}, and Mary B. Leonard, MD, MSCE^{1,2}

¹Division of Nephrology, Department of Pediatrics, The Children's Hospital of Philadelphia, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

²Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

³Department of Medicine, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

Abstract

Purpose—Chronic kidney disease (CKD) is a prevalent and important outcome and covariate in pharmacoepidemiology. The Health Improvement Network (THIN) in the U.K. represents a unique resource for population-based studies of CKD. We compiled a valid list of Read codes to identify subjects with moderate to advanced CKD.

Methods—A cross-sectional validation study was performed to identify codes that best define CKD stages 3–5. All subjects with at least one non-zero measure of serum creatinine after 1-Jan-2002 were included. Estimated glomerular filtration rate (eGFR) was calculated according to the Schwartz formula for subjects <18 years and the Modification of Diet in Renal Disease formula for subjects ≥18 years of age. CKD was defined as an eGFR <60 ml/min/1.73m² on at least two occasions, more than 90 days apart.

Results—The laboratory definition identified 230,426 subjects with CKD, for a period prevalence in 2008 of 4.56% (95% CI: 4.54, 4.58). A list of 45 Read codes was compiled which yielded a positive predictive value of 88.9% (95% CI: 88.7, 89.1), sensitivity of 48.8%, negative predictive value of 86.5%, and specificity of 98.2%. Of the 11.1% of subjects with a code who did not meet the laboratory definition, 83.6% had at least one eGFR <60. The most commonly used code was for CKD stage 3.

Conclusions—The proposed list of codes can be used to accurately identify CKD when serum creatinine data are limited. The most sensitive approach for the detection of CKD is to use this list to supplement creatinine measures.

Keywords

Chronic kidney disease (CKD); validity; database; epidemiology

Corresponding author: Michelle Denburg, MD, Division of Nephrology, The Children's Hospital of Philadelphia Main Building, Room 2143, Philadelphia, PA 19104, Phone: 215-590-2449, Fax: 215-590-3705, denburgm@email.chop.edu.

Conflict of Interest statement: We received the data from Cegedim who owns THIN, and co-authors Lewis and Haynes have participated in other research on THIN supported in part by Cegedim.

INTRODUCTION

Chronic kidney disease (CKD) is an increasingly prevalent condition with considerable morbidity and mortality. The medication utilization by this patient population and the modified dosing regimens required for moderate to advanced stages of renal insufficiency make CKD an important exposure, outcome and covariate in pharmacoepidemiologic research. Epidemiologic studies in children and adults with pre-dialysis CKD have been hampered by the absence of longitudinal, population-based data that include repeated measures of renal function, medications, and clinical outcomes.

The Health Improvement Network (THIN), a primary care medical records database in the United Kingdom, represents an under-utilized resource for studies of CKD, both as an exposure and outcome.¹ The system of primary care in the U.K. facilitates population-based studies because its coverage is extensive, and the general practitioner (GP) acts as a gatekeeper to patient care and has almost exclusive prescribing rights. THIN provides longitudinal records with data on demographics, growth parameters, medical diagnoses, laboratory values, prescriptions, and specialist referrals. The database also offers a robust sample of unexposed subjects that can be matched by year of birth, sex, and GP practice. As such, THIN has the potential to be a valuable tool for studying CKD.

A necessary first step to performing epidemiologic studies of CKD using THIN is to establish the validity and sensitivity of the diagnosis of CKD. While THIN contains laboratory data for some patient encounters, particularly in more recent years, diagnoses are also recorded in the database using Read codes. The validity of the numerous renal Read codes has not been previously addressed. Obtaining a valid list of codes for CKD will be useful in the absence of documented laboratory data, particularly for long-term follow-up of subjects diagnosed with CKD before the availability of laboratory data. The objective of this cross-sectional validation study was to identify the codes that best define moderate to advanced CKD (stages 3–5) within THIN, facilitating future longitudinal pediatric and adult studies in this high-risk population.

METHODS

Data source

THIN data collection from general practices was established in 2002 as a collaborative effort between InPractice Systems (provides the Vision software used in general practices) and EPIC (provides primary care data for medical research). THIN provides anonymous data derived from the daily records of participating practices and comprises data from more than seven million patients, of whom at least three million are actively registered. While prospective data collection for THIN started in 2002, the database includes practice data from the date a practice became computerized, which for some of the THIN practices dates back as far as 1987. The October 2008 version of the database was used in this analysis.

Read codes

A broad table of 973 Read codes potentially indicative of underlying renal disease was generated from look-up tables provided by THIN and included ambiguous codes, general codes that describe a non-specific impairment in renal function, and specific codes that refer to the stage of CKD or type of renal replacement therapy. The study evaluated the validity of a subset of 80 Read codes that were determined a priori to be most clinically indicative of moderate to advanced CKD, and were used at least 10 times in the database.

Subjects

Subjects were included if they had at least one non-zero measure of serum creatinine (sCr) after January 1, 2002. Since ages 90 or greater constitute direct identifiers, subjects who were greater than 90 years of age at any point during the study period were excluded as a privacy protection. Subjects without acceptable records for research (based on data checks performed by the data vendor) were also excluded. Children do not attain adult values for mean glomerular filtration rate (GFR) until approximately two years of age.^{2, 3} Therefore, the sCr measures performed in subjects who were less than two years of age at the time of measurement were excluded from the analysis.

Estimation of renal function from serum creatinine

For subjects less than 18 years of age at the time of sCr measurements, estimated glomerular filtration rate (eGFR) was calculated according to the Schwartz formula: $eGFR = [k \times \text{height (cm)}] / [\text{creatinine } (\mu\text{mol/l}) / 88.4]$ where k is a correction factor that depends on sex and age.⁴ A correction factor of 0.55 was used for girls less than 18 and boys less than 13 years of age, and a factor of 0.7 was used for adolescent boys ≥ 13 years of age. If the subject had a height recorded within 18 months of the sCr measure, the height within the nearest quarter to that of the sCr was used. If there were multiple heights available within this nearest quarter, or two heights within the same number of quarters from that of the sCr measure, the average of these heights was used. For sCr measures without an associated height within 18 months, the 50th percentile for age (at time of sCr measurement) and sex was used to impute height. This height imputation was performed using LMSgrowth software to access UK 1990 reference data based on the LMS method.^{5, 6} For subjects 18 years of age or older, eGFR was calculated from the sCr ($\mu\text{mol/l}$) according to the Modification of Diet in Renal Disease (MDRD) formula: $eGFR = 32,788 \times (\text{sCr})^{-1.154} \times (\text{age})^{-0.203} \times 0.742$ (if female).^{7–11} These calculated estimates of GFR were used to define our “gold standard” for diagnosing CKD.

Definition of CKD

The widely accepted 2002 National Kidney Foundation’s Kidney Disease Outcomes Quality Initiative (KDOQI) clinical practice guidelines defined CKD as persistent evidence of kidney damage or reduced GFR ($<60 \text{ ml/min/1.73m}^2$) for at least three months.¹² In accordance with the KDOQI definition and 5-stage classification, adopted in the U.K.,⁹ moderate to severe CKD (stages 3–5) was defined as an eGFR $<60 \text{ ml/min/1.73m}^2$ on at least two occasions, more than 90 days apart. To assess potential misclassification of acute kidney injury by the aforementioned definition, a sensitivity analysis was performed using a stricter definition of an eGFR $<60 \text{ ml/min/1.73m}^2$ on at least two occasions, more than 90 days apart, without an intervening or subsequently normal eGFR ($\geq 90 \text{ ml/min/1.73m}^2$). Read codes indicating normalization of kidney function were not found.

Descriptive statistics for continuous variables were reported as median and inter-quartile or total ranges for skewed data. The Wilcoxon rank sum test was used to compare distributions of eGFR. A two-sided p-value of <0.05 was considered statistically significant. Summary statistics (predictive values, sensitivity and specificity) for Read codes and exact binomial confidence intervals for proportions were generated. All analyses were performed using STATA 10.0 (Stata Corporation, College Station, TX).

The study was approved by the Institutional Review Boards of the University of Pennsylvania and the Children’s Hospital of Philadelphia.

RESULTS

Serum creatinine measures

Figure 1 demonstrates how the final 6,048,159 sCr measures used in the analysis were identified. There were 6,056,493 non-zero values of sCr after January 1, 2002 among 1,632,819 subjects with acceptable records for research who had neither died nor transferred out of their THIN practice prior to 2002. The 1,493 (0.02%) sCr measures performed in subjects who were less than two years of age at the time of measurement were excluded. All entries of laboratory values in THIN are assigned a flag indicative of the integrity of the record. 99.9% of the entries for sCr had acceptable records. There were 1,174 (0.02%) records with a flag indicating that either the day and/or month components of the date of measurement were imputed, and these values were excluded since the definition of CKD is based on timing between measures of compromised GFR. There were 2,915 (0.05%) records with a flag indicating that the units of measure were invalid. However, these sCr values were not excluded, because all 2,915 had no units of measure entered and a distribution [median 82, interquartile range (IQR) 70–96] similar to that among values recorded in $\mu\text{mol/l}$, the standard unit of measure in the U.K. (median 87, IQR 75–102).

Over 95% of the remaining sCr values were documented in $\mu\text{mol/l}$, 3.8% of values had no units, and 0.9% had units that were clearly typographical errors of the standard unit of measure (mmol/L, mol/L, and MicroU/L). The distribution of sCr among the values with these units of measure or no units (median 87, IQR 75–103) was similar to that among values in the standard $\mu\text{mol/l}$ (median 87, IQR 75–102). The 1554 (0.03%) values of sCr with other associated ambiguous or infrequently used units were excluded. Finally, 3984 (0.07%) values of sCr that were less than 20 $\mu\text{mol/l}$ (0.2 mg/dl) and 129 (0.002%) values that were greater than 1800 $\mu\text{mol/l}$ (20 mg/dl) were determined to fall outside the realm of biologic plausibility and were excluded.

Estimation of renal function from sCr

Of the 6,048,159 sCr measures included in the analysis, 5,963,104 (98.6%) were performed in subjects 18 years of age or older. The median eGFR, using the MDRD formula, was 71.1 ml/min/1.73m² (IQR 58.9–83.9).

Assignment of heights to sCr measures—There were 85,055 measures of sCr performed among 64,440 subjects who were between two and 18 years of age at the time of measurement. Table 1 shows the distribution of the difference in time between the sCr measures and their nearest respective heights in quarters. Over 43% of measurements had at least one height recorded within 18 months, and 65% of these sCr values had a height recorded within two quarters of the sCr measurement. For the 48,352 (57%) sCr measures for which there was no associated height within 18 months, height was imputed from the 50th percentile for sex and age (at time of sCr measurement).

Figure 2a shows the scatterplot of height by age at the time of sCr measurement and demonstrates that there were clearly implausible values among the actual heights recorded within 18 months of their respective sCr measurements. To address these erroneous heights, the LMS method was used to assign a sex- and age-based standard deviation score (SDS) to every height. The 539 actual height values with an SDS >+3 or <−4 were replaced with the 50th percentile values. The lower limit of −4 for the range of plausible standard deviation scores was chosen to account for the linear growth retardation that can occur in children with advanced CKD.^{13, 14} Figure 2b shows the revised scatterplot of height by age at the time of sCr measurement.

Using this combination of actual and imputed heights and the Schwartz formula to calculate eGFR for the 85,055 sCr measures in subjects less than 18 years of age, the median eGFR was 121.1 ml/min/1.73m² (IQR 106.4–139.4). To explore the sensitivity of eGFR to the height imputation, the distribution of eGFR using actual height values was compared to that using 50th percentile values, and they were found to be highly similar [median 120.4 (IQR 105.7–138.5) versus median 121.6 (IQR 107.1–139.8), respectively]. The statistical significance ($p < 0.0001$) of this clinically insignificant overestimation of eGFR can be attributed to the large sample size. There was no statistically significant difference found between the distributions of eGFR using actual versus imputed height among the small subset of eGFR values representing compromised renal function (<60 ml/min/1.73m²), for which the height imputation would have been most likely to overestimate actual height at the time of sCr measurement [median 38.3 (IQR 23.1–49.6, $n = 289$ actual) versus median 41.4 (IQR 27.7–52.9, $n = 150$ imputed), $p = 0.13$]. To further explore the sensitivity of eGFR to the height imputation, the distribution of eGFR using actual height values was compared to that using imputed 50th percentile height values for the 36,164 sCr values with a plausible height within 18 months of the sCr measurement and were found to be highly similar (median 120.4, IQR 105.7–138.5, versus median 120.0, IQR 105.3–137.6, respectively), and again even among the small subset of eGFR values representing compromised renal function (median 38.3, IQR 23.1–49.6, versus median 37.7 (IQR 23.0–49.8). The median actual height associated with these 36,164 sCr measures was 162.6 cm (IQR 153.1–169.5) compared to a median imputed height of 163.3 cm (IQR 156.5, 163.5). As expected the absolute difference in eGFR using actual versus imputed height was greater than 20 ml/min/1.73m² only for the 305 sCr measures associated with actual heights at the extremes of the growth curve. The median absolute value of actual height SDS for these sCr measures was 2.83 with a minimum of 1.61.

Definition of CKD

We identified 230,426 subjects with an eGFR <60 ml/min/1.73m² on at least two occasions, more than 90 days apart, meeting criteria for the definition of CKD. There was a female preponderance (63.5%). The median age of CKD subjects was 72 years (IQR 64.3–78.1). When the stricter definition of an eGFR <60 ml/min/1.73m² on at least two occasions, more than 90 days apart, without an intervening or subsequently normal eGFR (≥ 90) was applied, the number of subjects meeting criteria for CKD dropped by 1.5% to 227,061.

Validity of Read codes

Using the aforementioned CKD definition of an eGFR <60 ml/min/1.73m² on at least two occasions, more than 90 days apart, as a gold standard, the diagnostic test characteristics of each code in a set of 80 Read codes, determined a priori to be most clinically indicative of moderate to advanced CKD and used at least 10 times in the database, were evaluated. The denominator population for the diagnostic testing was comprised of the 998,443 subjects who had at least two sCr measures i.e. who had the opportunity to meet the gold standard definition of CKD. Figure 3 displays the positive predictive values with their respective 95% confidence intervals for each of the component codes. The sensitivity, specificity, and positive and negative predictive values as well as the numbers of true positives, false positives, true negatives and false negatives for each of the component codes are available in the Appendix. There were 45 codes with a positive predictive value of $\geq 84\%$ and for which the lower bound of the 95% confidence interval was $\geq 80\%$, and these codes are displayed in Table 2. The most commonly used Read code was that for “Chronic kidney disease stage 3.” The diagnostic test characteristics of a composite list of these 45 codes are shown in Table 3. The positive predictive value of this composite list was 89% (95% CI: 88.7, 89.1). Using the stricter definition of an eGFR <60 ml/min/1.73m² on at least two occasions, more than 90 days apart, without an intervening or subsequently normal eGFR to define CKD, its positive

predictive value was 88% (95% CI: 88.0, 88.4). We performed a stratified analysis based on age at the time of the first serum creatinine measure and found that while the specificity remained high in those under 18 years of age (99.6%), the positive predictive value of the composite code list decreased to 46.6%, and the sensitivity increased substantially to 76.4% (Table 3b). Of the 126,463 subjects with at least one of the 45 codes, 124,151 (98.2%) had at least one eGFR below 60 ml/min/1.73m². There were another 21 codes, of the 80 tested, for which the point estimate of the positive predictive value was ≥80%, but the lower bound of the 95% CI fell below 80%, and which were therefore, designated as indeterminate. As reflected in the width of the confidence intervals, these codes were used infrequently in our population of subjects with at least two sCr measures; all but four of the codes were used in fewer than 100 subjects.

The other renal Read codes assigned to the 117,778 subjects meeting the gold standard definition of CKD, but without an entry for any of the 80 Read codes tested, were explored. Only 34,961 (29.7%) of these subjects had an entry for one of the broader list of 973 Read codes potentially indicative of renal disease. The most frequently used of these codes are displayed in Table 4. The diagnostic performance characteristics of these codes were subsequently evaluated, and with the exception of four codes, all had positive predictive values below 80%. The codes with positive predictive values above 80% were: “Serum creatinine raised,” “Renal function tests abnormal,” “[D]Renal function test abnormal,” and “Renal function monitoring”). These codes were not added to our proposed list, because their predictive value was likely driven by the association of the codes with the sCr measurements used to define CKD rather than the provider entering the codes to indicate CKD.

Prevalence of CKD

Using as a denominator the 4,497,749 subjects with acceptable records for research who had not transferred out of THIN prior to 2002 and were alive as of 2008, and subtracting from the numerator those subjects with CKD who had died before 2008, the period prevalence of CKD as of 2008 was estimated to be 4.56% (95% CI 4.54, 4.58). The prevalence stratified on age was as follows: 0.003% for age <18, 0.17% for age 18–44, 3.72% for age 45–64, and 27.17% for age ≥65 years. Prevalence was 4.49% (95% CI 4.47, 4.51) using the stricter definition of CKD.

There were 133,810 subjects who had at least one entry during the study period for any code in our proposed list of 45 codes. 21,411 (16%) of these subjects did not meet criteria for the laboratory definition of CKD, and 2494 (1.9%) never had a valid sCr measure during the same period. Including these additional 21,411 subjects identified as having CKD by our proposed code list would increase the estimate of the period prevalence of CKD to 4.98% (95% CI 4.96, 5.00).

DISCUSSION

This study demonstrated that the THIN database comprises robust sCr data that can be used to identify a large population-based cohort of subjects with moderate to advanced CKD, and identified a list of 45 Read codes that can be used to accurately identify CKD in the absence of complete laboratory data with a positive predictive value of 89% (95% CI: 88.7, 89.1). The estimates of the period prevalence of CKD in our study of 4.56% (based on the gold standard laboratory definition) and 4.98% (based on this definition and our validated list of 45 codes) were consistent with U.K. prevalence estimates from other studies and data sources of 4.0–8.5%^{15–19} as well as with the estimated U.S. prevalence of CKD stages 3–4 from the National Health and Nutrition Examination Survey (NHANES) of 5.6% for NHANES III (1988–1994) and 8.0% for NHANES rounds 1999–2004.²⁰ The female

predominance of this CKD population was also consistent with other U.K. and U.S. reports.^{7, 18–20} Even including subjects identified by our proposed code list, the estimated period prevalence in our study was likely still an underestimate given that some subjects may not have had the requisite creatinine measures to make the classification of CKD.

Based on a survey of all renal units in the U.K. in 2004, there were an estimated 140,000 CKD patients under the care of nephrologists, indicating that as in the U.S., a substantial proportion of the CKD population, particularly patients with CKD stage 3, are cared for by general practitioners.²¹ In a study of eGFR reporting in a U.K. Primary Care Trust of 185,434 patients, only 3.7% of patients identified as having CKD stage 3–5 were under nephrology care versus 82.6% in primary care and 13.7% in other secondary care.¹⁸ Thus, large primary care databases with sCr data and diagnostic coding of CKD, such as THIN, are important resources for population-based studies of CKD.

This study has several important limitations. We based our “gold standard” definition of CKD on the widely accepted 5-stage classification of the 2002 KDOQI guidelines.¹² While the 90-day interval between measures of compromised eGFR is consistent with the KDOQI definition of chronicity and is designed to minimize misclassification of acute kidney injury, some degree of misclassification still remains, particularly for subjects with only two measures of compromised eGFR or for whom multiple measures were clustered in two time periods i.e. subjects with distinct episodes of acute kidney injury may have been misclassified as having CKD. To address this potential for misclassification, we performed a sensitivity analysis using a stricter CKD definition of an eGFR <60 ml/min/1.73m² on at least two occasions, more than 90 days apart, without an intervening or subsequently normal eGFR (≥ 90 ml/min/1.73m²), and the number of subjects meeting criteria for CKD dropped minimally by 1.5% with negligible impact on the results of the Read code validation. Furthermore, while some of these subjects may have had episodes of acute kidney injury that were misclassified, others may have had CKD and undergone renal transplantation with normalization of their eGFR. Therefore, to be inclusive and not omit renal transplant recipients, the stricter definition was not used as the gold standard.

There was also the potential for misclassification by our gold standard CKD definition secondary to incomplete or missing laboratory data. Subjects with true CKD may not have been identified as having CKD by this definition if they had only one eGFR <60 ml/min/1.73m² within the study period, but they had other measures of compromised eGFR that were recorded before 2002, not recorded in the absence of or problems with the electronic link to the laboratory, or were previously ≥ 60 but <90 ml/min/1.73m². This potential source of misclassification was highlighted by the proportion of false positives with an eGFR <60 ml/min/1.73m² recorded on at least one occasion. Over 83% of the subjects with at least one entry for a code in our validated list of 45 codes who did not meet criteria for the gold standard laboratory-based definition of CKD had at least one eGFR <60 ml/min/1.73m².

Using an estimated rather than a measured GFR raises another source of potential misclassification. Froissart et al. demonstrated considerable misclassification by both the MDRD and Cockcroft-Gault equations when compared to renal clearance of ⁵¹Cr-EDTA.²² However, the various methods available to measure GFR are costly, time-consuming and impracticable clinically, particularly for purposes of ongoing surveillance and especially in a primary care setting.²³ Therefore, since the primary objective of this study was to identify a valid list of Read codes for CKD, we decided to base the “gold standard” definition of CKD on the most widely recommended and used methods of estimating GFR from sCr since these are the methods employed by practitioners to diagnose and code CKD. Thus, while we may have misclassified a subject with an eGFR >60 ml/min/1.73m² as having CKD stage 3 by the MDRD formula, the GP would likely have done so as well.

The degree of misclassification of CKD was likely greater among certain subgroups in THIN. The MDRD formula involves multiplying by a correction factor for black race. The lack of availability of data on race in the database and consequent inability to apply this correction factor would lead to greater underestimation of renal function among black subjects. There is also a normal age-related decline in GFR with consequent underestimation of renal function in the elderly, particularly women.^{7, 18–20} Given that our CKD cohort included patients up to 90 years of age, some of these patients may have been misclassified as having CKD by the MDRD formula.

When estimating GFR in subjects less than 18 years of age at the time of sCr measurement, we were faced with the challenge of substantial missing height data. For greater than 50% of the sCr measures in these subjects, there was no associated height within 18 months, and height was imputed from the 50th percentile for sex and age. The highly similar distributions of eGFR using actual versus imputed height, even among the small subset of eGFR values representing compromised renal function, demonstrated that this imputation performed well. As anticipated the absolute difference in eGFR using actual versus imputed height was greater than 20 ml/min/1.73m² only for the minority of measures (0.8%) associated with actual heights at the extremes of the growth curve. Therefore, while the Schwartz formula is not particularly sensitive to a wide range of heights within the norms of the growth curve, the limitation of imputing height applies to very tall children in whom renal function may be underestimated, and children with profound short stature in whom renal function may be overestimated, the latter being of greater concern in a study of pediatric CKD. Thus, while our method of imputation generally performed well, the frequency of missing height data in children underscores the importance of a valid list of Read codes to identify or confirm CKD.

The proposed code list performed differently in children compared to adults. The decrease in positive predictive value despite an increase in specificity to >99% and an increase in sensitivity was likely driven by the considerable drop in the prevalence of CKD in subjects under 18 years of age. Predictive value is influenced by prevalence, and CKD is extremely rare in children, as corroborated by the estimated prevalence of CKD in THIN in subjects <18 years of age of 0.003%.

Interpretation of our data on the sensitivity of the Read codes requires careful consideration of the source cohort. We measured the sensitivity of Read codes among those subjects with recorded laboratory data. For these subjects, the GP may feel that recording the laboratory data is sufficient to identify the presence of CKD in the medical record and that entering a Read code for CKD is redundant. As such, it is possible that the sensitivity of the Read codes to identify CKD among subjects without laboratory data could be even higher than we have estimated. At least some proportion of the false negative population can also be attributed to the established and considerable under-recognition of CKD in the adult population^{19, 24}. Exploration of frequently used additional codes among the false negative subjects (Table 4) demonstrated that some proportion had codes, such as that for CKD stage 2 or acute renal failure, indicative of misclassification either by our “gold standard” criteria for CKD stages 3–5 or miscoding by the GP. Some proportion had codes for underlying renal etiologies of CKD, such as polycystic kidney disease, and some had codes indicative of CKD, but for which the positive predictive value was <80%. Our objective was to identify a list of Read codes that was predictive of CKD in the absence of laboratory data, not to use this list exclusively for the detection of CKD. Therefore, our final selection of Read codes was driven by maximizing positive predictive value rather than sensitivity as it has been shown that high positive predictive value and specificity are needed for valid epidemiologic studies. We do not advocate relying on this list of Read codes exclusively; the

most sensitive approach to identify CKD in epidemiologic studies would be to use these codes to augment the laboratory-based definition, particularly in the pediatric population.

CONCLUSION

In summary, the THIN database comprises a large population-representative cohort of subjects with moderate to severe CKD and repeated measures of renal function, and represents an under-utilized resource to perform longitudinal pharmacoepidemiologic studies in this high-risk patient population. This study identified a list of 45 Read codes that can be used to predict CKD in the absence of documented laboratory data, particularly for identifying subjects with diagnoses of CKD in the 1990's. The most sensitive approach for the detection of CKD is to use the validated code list to supplement the creatinine-based definition.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sponsors:

NIH grant T32-DK07006-35 (MR Denburg)

NIH grant K24-DK 076808 (MB Leonard and J Shults)

NIH grant K24-DK078228 (JD Lewis)

NIH grant UL1-RR024134 (K Haynes and JD Lewis)

References

1. Margolis DJ, Hofstad O, Feldman HI. Association between renal failure and foot ulcer or lower-extremity amputation in patients with diabetes. *Diabetes Care*. 2008; 31:1331–6. [PubMed: 18390800]
2. Atiyeh BA, Dabbagh SS, Gruskin AB. Evaluation of renal function during childhood. *Pediatr Rev*. 1996; 17:175–80. [PubMed: 8935916]
3. Heilbron DC, Holliday MA, al-Dahwi A, Kogan BA. Expressing glomerular filtration rate in children. *Pediatr Nephrol*. 1991; 5:5–11. [PubMed: 2025537]
4. Schwartz GJ, Haycock GB, Edelmann CM Jr, Spitzer A. A simple estimate of glomerular filtration rate in children derived from body length and plasma creatinine. *Pediatrics*. 1976; 58:259–63. [PubMed: 951142]
5. Cole TJ. The LMS method for constructing normalized growth standards. *Eur J Clin Nutr*. 1990; 44:45–60. [PubMed: 2354692]
6. Pan, H.; Cole, TJ. LMSgrowth, a Microsoft Excel add-in to access growth references based on the LMS method. Version 2.69. <http://www.healthforallchildren.co.uk/2010>
7. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann Intern Med*. 1999; 130:461–70. [PubMed: 10075613]
8. National Collaborating Centre for Chronic Conditions. Chronic kidney disease: national clinical guideline for early identification and management in adults in primary and secondary care. London: Royal College of Physicians; September. 2008
9. Archibald G, Bartlett W, Brown A, et al. UK Consensus Conference on Early Chronic Kidney Disease--6 and 7 February 2007. *Nephrol Dial Transplant*. 2007; 22:2455–7. [PubMed: 17557775]

10. Crowe E, Halpin D, Stevens P. Early identification and management of chronic kidney disease: summary of NICE guidance. *BMJ*. 2008; 337:a1530. [PubMed: 18824486]
11. <http://www.renal.org/whatwedo/InformationResources/CKDeGUIDE.aspx>. (Accessed at
12. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Am J Kidney Dis*. 2002; 39:S1–266. [PubMed: 11904577]
13. Fine RN. Etiology and treatment of growth retardation in children with chronic kidney disease and end-stage renal disease: a historical perspective. *Pediatr Nephrol*. 25:725–32. [PubMed: 20033735]
14. Mehls O, Wuhl E, Tonshoff B, Schaefer F, Nissel R, Haffner D. Growth hormone treatment in short children with chronic kidney disease. *Acta Paediatr*. 2008; 97:1159–64. [PubMed: 18624988]
15. [Accessed October 13, 2010] EastMidlandsRenalNetwork.org.uk. 2010. <http://www.emrn.org.uk/>
16. de Lusignan S, Chan T, Stevens P, et al. Identifying patients with chronic kidney disease from general practice computer records. *Fam Pract*. 2005; 22:234–41. [PubMed: 15814579]
17. Quinn MP, Rainey A, Cairns KJ, et al. The practical implications of using standardized estimation equations in calculating the prevalence of chronic kidney disease. *Nephrol Dial Transplant*. 2008; 23:542–8. [PubMed: 17890743]
18. Richards N, Harris K, Whitfield M, et al. The impact of population-based identification of chronic kidney disease using estimated glomerular filtration rate (eGFR) reporting. *Nephrol Dial Transplant*. 2008; 23:556–61. [PubMed: 18065790]
19. Stevens PE, O'Donoghue DJ, de Lusignan S, et al. Chronic kidney disease management in the United Kingdom: NEOERICA project results. *Kidney Int*. 2007; 72:92–9. [PubMed: 17440495]
20. Castro AF, Coresh J. CKD surveillance using laboratory data from the population-based National Health and Nutrition Examination Survey (NHANES). *Am J Kidney Dis*. 2009; 53:S46–55. [PubMed: 19231761]
21. Ahmad A, Roderick P, Ward M, et al. Current chronic kidney disease practice patterns in the UK: a national survey. *QJM*. 2006; 99:245–51. [PubMed: 16497848]
22. Froissart M, Rossert J, Jacquot C, Paillard M, Houillier P. Predictive performance of the modification of diet in renal disease and Cockcroft-Gault equations for estimating renal function. *J Am Soc Nephrol*. 2005; 16:763–73. [PubMed: 15659562]
23. Schwartz GJ, Furth SL. Glomerular filtration rate measurement and estimation in chronic kidney disease. *Pediatr Nephrol*. 2007; 22:1839–48. [PubMed: 17216261]
24. John R, Webb M, Young A, Stevens PE. Unreferred chronic kidney disease: a longitudinal study. *Am J Kidney Dis*. 2004; 43:825–35. [PubMed: 15112173]

Key points

- THIN comprises a large population-based cohort of subjects with moderate to severe CKD as well as repeated measures of renal function.
- This study identified a list of 45 Read codes that can be used to predict CKD in the absence of documented laboratory data.
- The most sensitive approach for the detection of CKD is to use the validated code list to supplement a creatinine-based definition.

Figure 1.
Identification of sCr measures for analysis.

Figure 2a. **Figure 2b.** **Figure 2.**

Figure 2a. Assigned height versus age at time of sCr measurement.

Figure 2b. Assigned height versus age at time of sCr measurement with actual height replaced with 50th percentile if height SDS $>+3$ or <-4 .

Figure 3.
Positive predictive values and 95% confidence intervals for 80 Read codes determined a priori to be most clinically indicative of moderate to advanced CKD.

Table 1

Distribution of the difference in time between serum creatinine (sCr) measures and their nearest respective heights for participants with heights recorded within 18 months of the sCr.

Quarters between sCr and nearest height	Number of sCr measures (%)
0	11516 (31)
1	7405 (20)
2	5017 (14)
3	4058 (11)
4	3417 (9)
5	2894 (8)
6	2396 (7)
Total	36703 (100)

Table 2

READ codes for which lower bound of 95% confidence interval of positive predictive value $\geq 80\%$.

Read code	Description	True positives	False positives	Positive predictive value
1Z12.00	Chronic kidney disease stage 3	99,578	11,767	0.894 (0.892,0.896)
K05..00	Chronic renal failure	8,910	603	0.937 (0.932,0.941)
1Z13.00	Chronic kidney disease stage 4	8,388	325	0.963 (0.959,0.967)
1Z1..00	Chronic renal impairment	4,558	526	0.897 (0.888,0.905)
9hE0.00	Except chronic kidney disease qual indic: Patient unsuitable*	3,175	303	0.913 (0.903,0.922)
K060.00	Renal impairment	3,052	370	0.892 (0.881,0.902)
66i..00	Chronic kidney disease monitoring	3,027	367	0.892 (0.881,0.902)
K06..00	Renal failure unspecified	2,281	327	0.875 (0.861,0.887)
1Z14.00	Chronic kidney disease stage 5	1,712	88	0.951 (0.94,0.961)
9Ot0.00	Chronic kidney disease monitoring first letter	1,595	254	0.863 (0.846,0.878)
K08..00	Impaired renal function disorder	1,398	132	0.914 (0.899,0.927)
6AA..00	Chronic kidney disease annual review	1,212	107	0.919 (0.903,0.933)
K060.11	Impaired renal function	949	86	0.917 (0.898,0.933)
1Z1C.00	Chronic kidney disease stage 3 without proteinuria	720	74	0.907 (0.884,0.926)
9Ot..00	Chronic kidney disease monitoring administration	715	126	0.85 (0.824,0.874)
9hE1.00	Exc chronic kidney disease quality indicators: Inform dissen *	563	73	0.885 (0.858,0.909)
K050.00	End stage renal failure	491	43	0.919 (0.893,0.941)
7L1A200	Haemodialysis NEC	463	48	0.906 (0.877,0.93)
7L1A.11	Dialysis for renal failure	391	27	0.935 (0.907,0.957)
9hE..00	Exception reporting: chronic kidney disease quality indicato *	299	57	0.84 (0.798,0.876)
7A60100	Creation of arteriovenous fistula NEC	251	13	0.951 (0.917,0.974)
7L1A100	Peritoneal dialysis	250	18	0.933 (0.896,0.96)
K08z.00	Impaired renal function disorder NOS	250	36	0.874 (0.83,0.91)
K06..11	Uraemia NOS	197	32	0.86 (0.808,0.902)
14V2.00	H/O: renal dialysis	148	14	0.914 (0.859,0.952)
4519.00	Deteriorating renal function	147	10	0.936 (0.886,0.969)
D215000	Anaemia secondary to chronic renal failure	141	8	0.946 (0.897,0.977)
D215.00	Anaemia secondary to renal failure	125	9	0.933 (0.876,0.969)
7L1B000	Insertion of ambulatory peritoneal dialysis catheter	111	2	0.982 (0.938,0.998)
1Z1B.00	Chronic kidney disease stage 3 with proteinuria	106	3	0.972 (0.922,0.994)
1Z15.00	Chronic kidney disease stage 3A	71	6	0.922 (0.838,0.971)
8L50.00	Renal transplant planned	69	7	0.908 (0.819,0.962)
7L1A000	Renal dialysis	66	8	0.892 (0.798,0.952)
K05..11	Chronic uraemia	63	2	0.969 (0.893,0.996)
K0D..00	End-stage renal disease	57	2	0.966 (0.883,0.996)
G22..11	Nephrosclerosis	46	3	0.939 (0.831,0.987)

Read code	Description	True positives	False positives	Positive predictive value
ZV45100	[V]Renal dialysis status	44	3	0.936 (0.825,0.987)
7L1B100	Removal of ambulatory peritoneal dialysis catheter	38	3	0.927 (0.801,0.985)
1Z16.00	Chronic kidney disease stage 3B	37	0	1 (0.905,1)
1Z1E.00	Chronic kidney disease stage 3A without proteinuria	35	0	1 (0.9,1)
1Z1J.00	Chronic kidney disease stage 4 without proteinuria	35	1	0.972 (0.855,0.999)
1Z1H.00	Chronic kidney disease stage 4 with proteinuria	29	0	1 (0.881,1)
SP08300	Kidney transplant failure and rejection	25	1	0.962 (0.804,0.999)
1Z1G.00	Chronic kidney disease stage 3B without proteinuria	18	0	1 (0.815,1)
SP05613	[X] Peritoneal dialysis associated peritonitis	18	0	1 (0.815,1)

* Chronic kidney disease exception reporting codes; description abbreviated by data vendor

Table 3a

Diagnostic test characteristics of composite list of 45 Read codes.

	Gold standard definition of CKD	Stricter definition of CKD
Sensitivity (%)	48.8	49.1
95% CI	48.6, 49.0	48.9, 49.3
Specificity (%)	98.2	98.1
95% CI	98.1, 98.2	98.0, 98.1
PPV (%)	88.9	88.2
95% CI	88.7, 89.1	88.0, 88.4
NPV (%)	86.5	86.8
95% CI	86.4, 86.5	86.7, 86.8

PPV – Positive predictive value

NPV – Negative predictive value

Table 3b

Diagnostic test characteristics of composite list of 45 Read codes using the gold standard definition of CKD, stratified on age at first serum creatinine.

	<18 years of age	≥18 years of age
Sensitivity (%)	76.4	48.8
95% CI	64.9, 85.6	48.6, 49.0
Specificity (%)	99.6	98.1
95% CI	99.5, 99.7	98.1, 98.2
PPV (%)	46.6	88.9
95% CI	37.4, 56.0	88.7, 89.1
NPV (%)	99.9	86.2
95% CI	99.8, 99.9	86.1, 86.3

PPV – Positive predictive value

NPV – Negative predictive value

Table 4

Other READ codes used ≥ 100 times among subjects meeting criteria for CKD but without an entry from the previously tested list of 80 READ codes.

Read code	Description	Frequency*
44J5.00	Renal profile	100827
44JB.00	Urea and electrolytes	6543
1Z11.00	Chronic kidney disease stage 2	3043
9N1m.00	Seen in nephrology clinic	2606
9N1m.11	Seen in renal clinic	1744
R110.00	[D]Proteinuria	1630
451..00	Renal function tests	1243
44J..12	Urea and electrolytes	1006
K04..00	Acute renal failure	916
4678.00	Proteinuria	822
44J7.00	Albumin/creatinine ratio	786
44J3.00	Serum creatinine	614
451E.00	GFR calculated abbreviated MDRD	537
8A6..00	Renal function monitoring	493
K11..00	Hydronephrosis	407
451F.00	Glomerular filtration rate	396
8H4L.00	Referred to nephrologist	342
44J3300	Serum creatinine raised	300
R144.00	[D]Renal function test abnormal	299
7B01.00	Total nephrectomy	295
4676.00	Urine protein test = +++	284
K101.00	Acute pyelonephritis	284
ZL9AO00	Seen by nephrologist	246
467..00	Urine protein test	209
K01..00	Nephrotic syndrome	193
K18..00	Urethral stricture	192
451A.00	Estimated creatinine clearance	184
C10FL00	Type 2 diabetes mellitus with persistent proteinuria	141
K10y000	Pyelonephritis unspecified	130
1Z10.00	Chronic kidney disease stage 1	128
4512.00	Renal function tests abnormal	128
4674.00	Urine protein test = +	121
ZL5AN00	Referral to nephrologist	120
4677.00	Urine protein test = ++++	111
PD11.00	Polycystic kidney disease	109
8A6..11	Kidney function monitoring	101

Read code	Description	Frequency*
K132.00	Acquired cyst of kidney	100

* Refers to the number of times the code was used (not to the number of subjects).