

Predicting Hospitalizations for Patients With Chronic Kidney Disease

Steph Karpinski, MS; Scott Sibbel, PhD, MPH; Kathryn Gray, MA; Adam G. Walker, PhD; Jiacong Luo, MD; Carey Colson, MBA; Juliana Stebbins, MPH; and Steven M. Brunelli, MD, MSCE

Approximately 15% of individuals living in the United States have chronic kidney disease (CKD).¹ Those with CKD are twice as likely to be hospitalized as individuals in the general population.^{2,3} In those with CKD, hospitalization is associated with greater morbidity and mortality, including but not limited to progression to end-stage kidney failure.⁴ Hospitalizations also account for disproportionate resource utilization among patients with CKD; for example, Medicare expenditures for inpatient costs among patients with CKD exceeded \$24.8 billion in 2019 alone.⁵

In the United States, recent federal initiatives have increased focus on value-based care among patient populations with CKD.⁶ Under these managed care programs, clinicians and health care providers are incentivized for improving health outcomes and reducing costs for Medicare beneficiaries with a diagnosis of CKD. Similar programs exist for patients with insurance other than Medicare. Reducing hospitalization rates is an important goal because hospital admissions are associated with poorer clinical outcomes and increased medical costs in CKD. By extension, algorithms that enable a priori identification of patients at high near-term risk of being hospitalized and permit opportunity for targeted preventive care are of great importance. Commonly, insurers share historical medical claims information for attributed patients with health care providers as a part of managed care arrangements. Historical laboratory measures and electronic health records (EHRs) are shared in some instances but are not usually available for all patients. Because risk predictions cannot be made for patients for whom requisite predictor data are missing, a model based on claims data allows for predictions to be made on the greatest number of patients. Here, we report the development of an algorithm that uses historical medical claims to predict risk of hospitalization in the next 90 days among patients with CKD.

METHODS

Data and Patients

This model was developed and validated using historical, deidentified Medicare Part A and Part B claims from calendar years 2017 and 2018. According to 45 Code of Federal Regulations part 46 from HHS, this

ABSTRACT

OBJECTIVES: Patients with chronic kidney disease (CKD) are at higher risk of being admitted to the hospital than the general population. Hospitalizations in patients with CKD are associated with higher medical costs and increased morbidity and mortality. Identification of patients with CKD who are at greatest risk of hospitalization may hold promise to improve clinical outcomes and enable judicious allocation of health care resources.

STUDY DESIGN: Retrospective, observational cohort study.

METHODS: Medicare Part A and Part B claims from calendar years 2017 and 2018 from 50,000 unique patients with a diagnosis of stage 3 to 5 CKD were used for this study. Data were split into training ($n=40,000$) and test ($n=10,000$) sets. A variety of model types were built to predict all-cause hospitalization within 90 days.

RESULTS: The final model was a gradient-boosting machine with 399 input terms. The model demonstrated good ability to discriminate (area under the curve [AUC] for the receiver operating characteristic curve = 0.73), which was stable when tested in the test set (AUC = 0.73). The positive predictive value in the test set was 0.306, 0.240, and 0.216 at the 10%, 20%, and 30% thresholds, respectively. The sensitivity in the test set was 0.288, 0.453, and 0.609 at the 10%, 20%, and 30% thresholds, respectively.

CONCLUSIONS: We developed an algorithm that uses medical claims to identify Medicare patients with CKD stages 3 to 5 who are at highest risk of being hospitalized in the near term. This algorithm could be used as a decision support tool for clinical programs focusing on management of patient populations with CKD.

Am J Manag Care. 2023;29(9):e262-e266. doi:10.37765/ajmc.2023.89428

study was deemed exempt from institutional review board or ethics committee approval. We adhered to the Declaration of Helsinki, and informed consent was not required.

We considered a historical prediction date of July 1, 2018. Eligible patients were those who, as of that date, were enrolled in Medicare Part A and B, were 18 years or older, had a diagnosis of CKD stages 3 to 5 within the predictor data window (*International Classification of Diseases, Tenth Revision* codes N18.3, N18.4, N18.5 from July 1, 2017, to March 30, 2018), and had no prior claims for dialysis. If multiple CKD stages were indicated in the baseline claims, we assigned the patient the greatest of these CKD stages. To reduce computational time, we randomly selected 50,000 eligible patients and split them into training (80%; n = 40,000 patients) and test (20%; n = 10,000 patients) sets.

Model Design and Development

The model was designed to predict all-cause hospitalization in the 90-day period immediately following the historical prediction date (July 2, 2018, through September 30, 2018). Hospitalization for any cause during this period was considered as a binary outcome. Demographic information, diagnosis codes, and procedure codes abstracted from medical claims from the prior year were used as predictors (**Figure 1**). To simulate a real-world claims run-out period (the time between when medical service is rendered and the claim is fully processed and available for use), claims from the 90 days immediately preceding prediction (April 1, 2018, through June 30, 2018) were suppressed and not used for defining predictors.

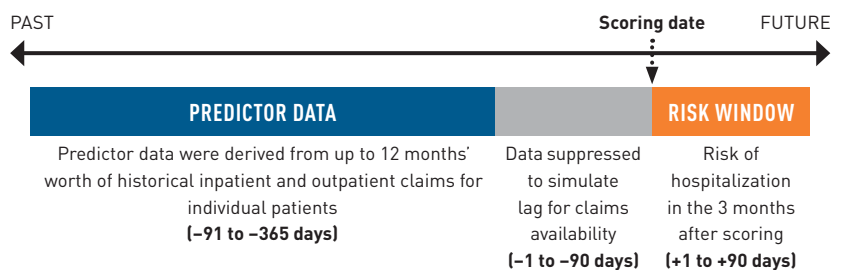
All data analysis and modeling were performed in R version 4.1.0 (R Project for Statistical Computing), primarily using the caret package.⁷ We tested various candidate models using a variety of model algorithms, including logistic regression, artificial neural networks, random forests, classification trees, multivariate adaptive regression splines, and support vector machines. After preprocessing the data, cross-validated candidate models were tuned in an iterative process using area under the curve (AUC) of the receiver operating characteristic curve as the representative model performance metric. In addition to AUC, positive predictive value (PPV) and sensitivity across a variety of thresholds as well as stability of model performance parameters between training and test data sets were taken into consideration when selecting the final model.

TAKEAWAY POINTS

Hospitalizations in patients with chronic kidney disease (CKD) are associated with worse clinical outcomes and higher costs. We developed an algorithm to predict which patients with CKD are at greatest risk of being hospitalized within 90 days.

- ▶ The algorithm requires only information from medical claims to make predictions and is therefore well suited for use within managed care environments.
- ▶ This algorithm could be used as a decision support tool for clinical programs focusing on management of patient populations with CKD.

FIGURE 1. Predictive Model Design^a



^aA historical scoring date was chosen as July 1, 2018. Outcomes were considered for the 90-day risk window immediately following; hospitalization from any cause (yes/no) was considered as a binary response variable. Predictor data were abstracted from historical claims data from up to 1 year preceding the scoring date; of these, we suppressed claims from the most recent 90 days prior to scoring to simulate claims run-out period, meaning that functionally, predictors were considered for days -91 to -365.

TABLE 1. Characteristics of Patients in Model Training and Test Data Sets

Characteristic	Training set (n = 40,000)	Test set (n = 10,000)
Age in years, mean (SD)	76.2 (10.5)	76.3 (10.7)
Female, n (%)	20,614 (51.5%)	5152 (51.5%)
Race/ethnicity, n (%)		
Black	5013 (12.5%)	1176 (11.8%)
White	32,496 (81.2%)	8210 (82.1%)
Asian	609 (1.5%)	149 (1.5%)
Hispanic	668 (1.7%)	157 (1.6%)
Other/unknown	1214 (3.0%)	308 (3.1%)
CKD stage, n (%)		
3	32,938 (82.3%)	8225 (82.2%)
4	6418 (16.0%)	1589 (15.9%)
5	644 (1.6%)	186 (1.9%)
Hospitalized during the risk window, n (%)	4174 (10.4%)	1065 (10.7%)

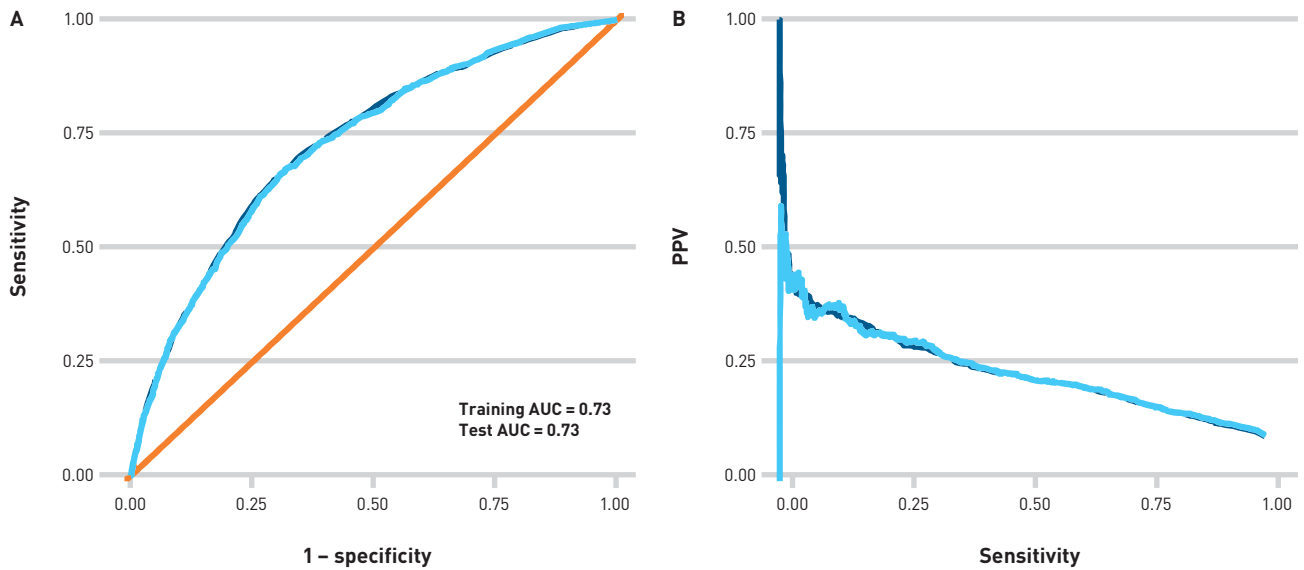
CKD, chronic kidney disease.

RESULTS

Patients

Patient characteristics at the time of the simulated scoring date are shown in **Table 1**. In the model training set, the mean age was 76 years, 51% were women, 81% were White, 12.5% were Black, and

FIGURE 2. Operating Characteristics of the Final Model^a



AUC, area under the curve; PPV, positive predictive value.

^aDepicted are the receiver operating characteristic curve (A) and precision-recall curve (B) in the training set (dark blue) and test set (light blue).

TABLE 2. Model Performance at Various Thresholds in the Training and Test Data Sets

High risk threshold	Training set (n = 40,000)	Test set (n = 10,000)
Top 10% of patients by risk score		
Sensitivity	0.285	0.288
Specificity	0.921	0.922
PPV	0.296	0.306
NPV	0.917	0.916
Top 20% of patients by risk score		
Sensitivity	0.466	0.453
Specificity	0.828	0.829
PPV	0.240	0.240
NPV	0.930	0.927
Top 30% of patients by risk score		
Sensitivity	0.613	0.609
Specificity	0.735	0.736
PPV	0.212	0.216
NPV	0.942	0.940

NPV, negative predictive value; PPV, positive predictive value.

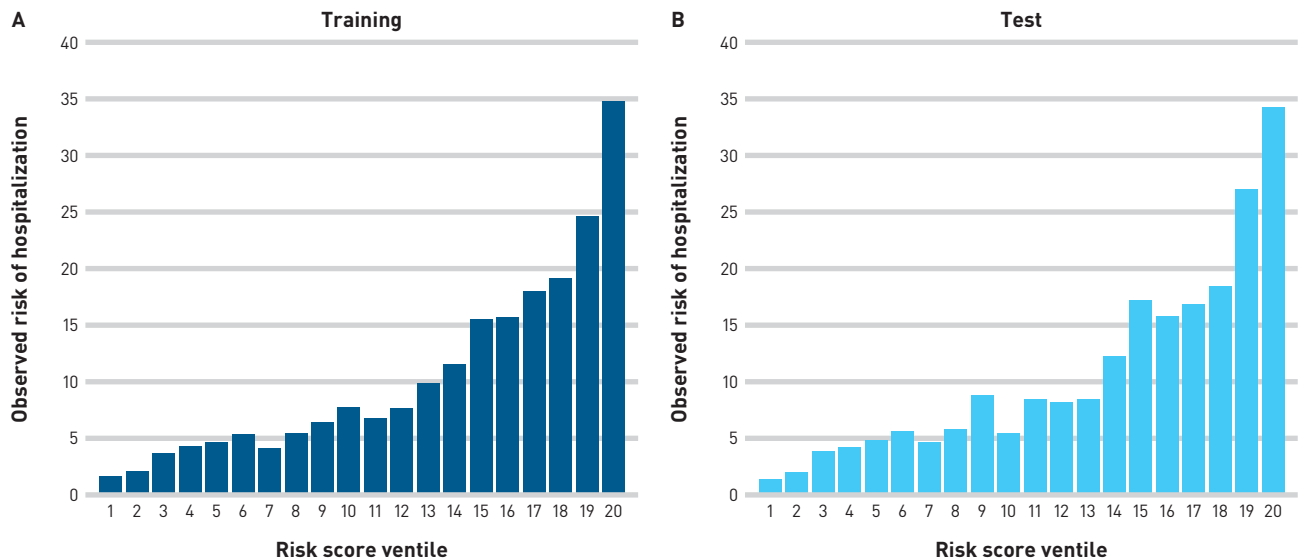
the vast majority of patients had CKD stage 3 (82%). Approximately 10% of these patients were hospitalized during the risk window. Characteristics of patients in the test set were nearly identical.

Modeling Results

The model with the best and most consistent performance was a gradient-boosting machine learning algorithm. There were 399 input terms selected into the model, which represented 147 unique clinical constructs. The model demonstrated good ability to discriminate (AUC = 0.73), which was stable when evaluated in the test set (AUC = 0.73) (Figure 2 [A]).

Figure 2 (B) contains precision-recall curves for the training and test sets, and Table 2 displays PPV and sensitivity at various thresholds. The PPV in the test set was 0.306, 0.240, and 0.216 at the 10%, 20%, and 30% thresholds, respectively. The sensitivity in the test set was 0.288, 0.453, and 0.609 at the 10%, 20%, and 30% thresholds, respectively. Among the subset of patients with stage 3 CKD, the model sensitivity in the test set was 0.282, 0.465, and 0.612 at the 10%, 20%, and 30% thresholds, respectively. Among the subset of patients with stage 4 or 5 CKD, the model sensitivity in the test set was 0.246, 0.426, and 0.544 at the 10%, 20%, and 30% thresholds, respectively.

Model calibration is shown in Figure 3 for the training and test data sets and is depicted as hospitalization risk by ventile of risk score. The observed hospitalization rate was approximately 35% in the top risk score ventile vs 1% in the lowest ventile, representing a 35-fold risk gradient.

FIGURE 3. Calibration Curves for the Final Model in the Training and Test Sets*

*Plotted are the observed hospitalization rates among patients representing ventiles of risk score. Risk score ventile 1 represents the lowest 5% of model risk scores, whereas risk score ventile 20 represents the highest 5% of model risk scores.

DISCUSSION

We developed a predictive model to stratify Medicare patients with CKD by near-term risk of all-cause hospitalization. This model utilizes only information available from medical claims, meaning that predictions can be made for the vast majority of patients in CKD health management programs, including for those with varying degrees of interaction with a nephrologist and those whose care is split across health systems. This claims-based approach to prediction is well suited to population health management.

Our model showed an overall good ability to discriminate patients with CKD who will be hospitalized, with an AUC of 0.73. This is equivalent to the discriminant performance of claims-based hospitalization risk models in the general population.^{8,9} Based on the model performance we observed in the test data set, selecting the top 10% to 30% of patients with CKD at the highest predicted risk would account for upward of 61% of all admissions in the population. In turn, this enables the efficient targeting of hospitalization avoidance efforts to patients who would disproportionately benefit.

Our goal was to develop a model to predict hospitalizations among patients with CKD in a managed care environment. Other published studies describe models developed to predict hospitalizations.⁸⁻¹⁷ However, many of these models were developed to predict cause-specific hospitalizations (eg, hospitalizations due to heart failure) or for other specific patient populations without CKD (eg, older adult home care or pediatric patients). Compared with the general population, overall rates of hospitalization are higher for patients with CKD across nearly all admission types; therefore, estimating

cause-specific hospitalization risk seemed overly narrow. Moreover, patients with CKD have a unique health profile relative to both the general population and other patient groups. Thus, existing models were not likely to generalize to our intended use case, and a CKD-specific model to predict hospitalizations was sought. Although several predictive models for patients with CKD have been published, most predict risk of worsening kidney disease or kidney failure.¹⁸⁻²⁰ To our knowledge, this is the first model reported in the literature developed to predict hospitalizations specifically among patients with CKD.

For this model, only information derived from medical claims was used as predictors. Findings from a recently published study demonstrated that combining claims with other types of predictors, such as medications and EHR data, improved model performance for predicting hospitalizations.⁸ Although better predictions are always desirable, our model was designed to be broadly applicable for use in a traditional managed care program. Whereas claims are usually available to health care providers for all attributed patients, medication information and EHRs are not. In turn, this would preclude use of a model that predicts on the basis of both claims and EHR elements or at least render it useful in risk stratifying only a small subset of patients, limiting its use for population health management.

Limitations

The limitations of this study should be acknowledged. First, the model was designed specifically for use among patient populations

with diagnosed CKD stages 3 to 5 and is not likely to generalize to other patient populations, including those with unrecognized CKD. This model was developed using claims data from patients with Medicare insurance. Because Medicare patients tend to be older, model performance may be different if applied to patients with CKD with other insurance types, such as younger patients who are commercially insured. Moreover, data used in this model predated the COVID-19 pandemic. Primary and secondary predictors related to COVID-19 were not able to be considered.

CONCLUSIONS

We have developed and validated an algorithm to predict all-cause hospitalization over a 90-day period for patients with CKD. This model utilizes only claims data as predictors and would therefore be suitable for use as a tool for population health management and value-based care programs. ■

Author Affiliations: DaVita Clinical Research (SK, SS, KG, AGW, JL, CC, SMB), Minneapolis, MN; DaVita, Inc (JS), Denver, CO; DaVita Integrated Kidney Care (JS), Denver, CO.

Source of Funding: None.

Author Disclosures: Ms Stebbins is employed by Davita, Inc and DaVita Integrated Kidney Care. All other authors are employed by DaVita Clinical Research. Ms Karpinski, Dr Sibbel, Ms Gray, Dr Walker, Dr Luo, Ms Stebbins, and Dr Brunelli report owning stock in DaVita, Inc.

Authorship Information: Concept and design (SK, SS, KG, AGW, JL, CC, JS, SMB); acquisition of data (KG, CC, SMB); analysis and interpretation of data (SK, KG, AGW, JL, JS, SMB); drafting of the manuscript (SS, AGW, CC, SMB); critical revision of the manuscript for important intellectual content (SK, SS, KG, AGW, JL, SMB); statistical analysis (SK, AGW, SMB); administrative, technical, or logistic support (JS); and supervision (SS).

Address Correspondence to: Steven M. Brunelli, MD, MSCE, Davita, Inc, 825 S 8th St, Ste 300, Minneapolis, MN 55404. Email: Steven.Brunelli@DaVita.com.

REFERENCES

- United States Renal Data System. CKD in the general population. In: *2021 USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States*. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; 2021. Accessed July 1, 2022. <https://usrds-adr.niddk.nih.gov/2021/chronic-kidney-disease/1-ckd-in-the-general-population>
- Schrauben SJ, Chen HY, Lin E, et al; CRIC Study Investigators. Hospitalizations among adults with chronic kidney disease in the United States: a cohort study. *PLoS Med*. 2020;17(12):e1003470. doi:10.1371/journal.pmed.1003470
- United States Renal Data System. Morbidity and mortality in patients with CKD. In: *2021 USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States*. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; 2021. Accessed July 1, 2022. <https://usrds-adr.niddk.nih.gov/2021/chronic-kidney-disease/3-morbidity-and-mortality-in-patients-with-ckd>
- Srivastava A, Cai X, Mehta R, et al; CRIC Study Investigators. Hospitalization trajectories and risks of ESKD and death in individuals with CKD. *Kidney Int Rep*. 2021;6(6):1592-1602. doi:10.1016/j.ekir.2021.03.883
- United States Renal Data System. Healthcare expenditures for persons with CKD. In: *2021 USRDS Annual Data Report: Epidemiology of Kidney Disease in the United States*. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases; 2021. Accessed July 1, 2022. <https://usrds-adr.niddk.nih.gov/2021/chronic-kidney-disease/6-healthcare-expenditures-for-persons-with-ckd>
- Kidney Care Choices (KCC) model. CMS. Updated January 20, 2023. Accessed April 5, 2022. <https://innovation.cms.gov/innovation-models/kidney-care-choices-kcc-model>
- Kuhn M. Caret: classification and regression training: version 6.0-91. Comprehensive R Archive Network. 2022. Accessed July 1, 2021. <https://topepo.github.io/caret/>
- Morawski K, Dvorkis Y, Monsen CB. Predicting hospitalizations from electronic health record data. *Am J Manag Care*. 2020;26(1):e7-e13. doi:10.37765/ajmc.2020.42147
- van der Galien OP, Hoekstra RC, Gürgöze MT, et al. Prediction of long-term hospitalisation and all-cause mortality in patients with chronic heart failure on Dutch claims data: a machine learning approach. *BMC Med Inform Decis Mak*. 2021;21(1):303. doi:10.1186/s12911-021-01657-w
- Álvarez-García J, Ferrero-Gregori A, Puig T, et al; Investigators of the Spanish Heart Failure Network (REDINSCOR). A simple validated method for predicting the risk of hospitalization for worsening of heart failure in ambulatory patients: the Redin-SCORE. *Eur J Heart Fail*. 2015;17(8):818-827. doi:10.1002/ehf.287
- Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open*. 2020;3(1):e1918962. doi:10.1001/jamanetworkopen.2019.18962
- Holloway J, Neely C, Yuan X, et al. Evaluating the performance of a predictive modeling approach to identifying members at high-risk of hospitalization. *J Med Econ*. 2020;23(3):228-234. doi:10.1080/13696998.2019.1666854
- Maltenfort MG, Chen Y, Forrest CB. Prediction of 30-day pediatric unplanned hospitalizations using the Johns Hopkins Adjusted Clinical Groups risk adjustment system. *PLoS One*. 2019;14(8):e0221233. doi:10.1371/journal.pone.0221233
- Monsen KA, Swanberg HL, Oancea SC, Westra BL. Exploring the value of clinical data standards to predict hospitalization of home care patients. *Appl Clin Inform*. 2012;3(4):419-436. doi:10.4338/ACI-2012-05-RA-0016
- Pauly V, Mendizabal H, Gentile S, Auquier P, Boyer L. Predictive risk score for unplanned 30-day rehospitalizations in the French universal health care system based on a medico-administrative database. *PLoS One*. 2019;14(3):e0210714. doi:10.1371/journal.pone.0210714
- Vaughn DA, van Deen WK, Kerr WT, et al. Using insurance claims to predict and improve hospitalizations and biologics use in members with inflammatory bowel diseases. *J Biomed Inform*. 2018;81:93-101. doi:10.1016/j.jbi.2018.03.015
- Verhoeff M, de Groot J, Burgers JS, van Munster BC. Development and internal validation of prediction models for future hospital care utilization by patients with multimorbidity using electronic health record data. *PLoS One*. 2022;17(3):e0260829. doi:10.1371/journal.pone.0260829
- Tangri N, Inker LA, Hiebert B, et al. A dynamic predictive model for progression of CKD. *Am J Kidney Dis*. 2017;69(4):514-520. doi:10.1053/j.ajkd.2016.07.030
- Tangri N, Stevens LA, Griffith J, et al. A predictive model for progression of chronic kidney disease to kidney failure. *JAMA*. 2011;305(15):1553-1559. doi:10.1001/jama.2011.451
- Zacharias HU, Altenbuchinger M, Schultheiss UT, et al; GCKD Investigators. A predictive model for progression of CKD to kidney failure based on routine laboratory tests. *Am J Kidney Dis*. 2022;79(2):217-230.e1. doi:10.1053/j.ajkd.2021.05.018

Visit ajmc.com/link/89428 to download PDF