# Identifying Patients with Rare Disease Using Electronic Health Record Data: The Kaiser Permanente Southern California Membranous Nephropathy Cohort

Amy Z Sun, MD[1]; Yu-Hsiang Shu, PhD[2]; Teresa N Harrison, SM[2]; Aviv Hever, MD[3]; Steven J Jacobsen, MD, PhD[2]; Michelle M O'Shaughnessy, MD[4]; John J Sim, MD[1]

Perm J 2020;24:19.126

E-pub: 02/07/2020

https://doi.org/10.7812/TPP/19.126

## ABSTRACT

**Introduction:** Developing a reliable means to identify and study real-world populations of patients with membranous nephropathy (MN) using electronic health records (EHRs) would help advance glomerular disease research. Identifying MN cases using EHRs is limited by the need for manual reviews of biopsy reports.

**Objective:** To evaluate the accuracy of identifying patients with biopsy-proven MN using the EHR in a large, diverse population of an integrated health system.

**Methods:** A retrospective cohort study was performed between June 28, 1999, and June 25, 2015, among patients with kidney biopsy results (N = 4723), which were manually reviewed and designated as MN or non-MN. The sensitivity, specificity, and positive predictive value (PPV) of International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes were determined using 2 approaches: 1) clinical (MN-specific codes 581.1, 582.1, or 583.1) and 2) agnostic/data-derived (codes selected from supervised learning at the highest predictive performance).

**Results:** One year after biopsy, the sensitivity and specificity of an MN diagnosis were 86% and 76%, respectively, but the PPV was 26%. The data-driven approach detected that using only 2 codes (581.1 or 583.1) improved specificity to 94% and PPV to 58%, with a small decrease in sensitivity to 83%. When any code was reported at least 3 times, specificity was 98%; PPV, 78%; and sensitivity, 64%.

**Discussion:** Our findings suggest that ICD-9 diagnosis codes might be a convenient tool to identify patients with MN using EHR and/or administrative claims information. Codes selected from supervised learning achieved better overall performance, suggesting the potential of developing data-driven methods.

## INTRODUCTION

Membranous nephropathy (MN) is among the most common glomerular diseases in adults and is associated with substantial morbidity, mortality, and health care expenditure.[1,2] It is challenging to study large-scale populations of patients with MN because tools (eg, laboratory tests or diagnosis codes) to quickly identify cases of MN from electronic health record (EHR) or administrative/claims datasets are lacking. Indeed, the diagnosis of MN requires a kidney biopsy, which is the gold-standard diagnostic test for MN. However, manual review of large volumes of kidney biopsy reports is impeded by resource limitations and/or the unavailability of biopsy reports in structured research datasets. As a result, knowledge of MN epidemiology largely derives from single-center case series or resource-intensive cohort studies, with uncertain generalizability of findings to the overall population.[3-5]

The increasing prevalence of health systems capturing information through the EHR provides more opportunities to study rare disease populations such as glomerular disease and may help improve management strategies. Building an effective means to recognize and follow these populations within the backdrop of a real-world environment would provide insights into the natural history of disease and would help determine treatment. We postulate that developing a reliable means to identify and study real-world populations of patients with MN using existing administrative and clinical information would represent a major advancement in glomerular disease research.[6-8] Accordingly, we sought to develop an approach to readily and reliably identify patients with biopsy-proven MN in a large integrated health system using International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes.

## METHODS

A retrospective cohort study was performed in Kaiser Permanente Southern California (KPSC), an integrated health system inclusive of more than 4.5 million members, 14 medical centers, and more than 200 satellite clinics. The study population was derived from 4723 patients who underwent a native kidney biopsy procedure between June 28, 1999, and June 25, 2015, and had at least 1 year of continuous follow-up.[9] A manual chart review of biopsy reports for all patients was performed to identify the different cases of glomerular disease including MN.[9] Secondary MN was excluded with the exception of MN associated with systemic lupus erythematosus, which was included in the study cohort. On the basis of the chart reviews, patients were categorized as having biopsy-proven MN or non-MN. This retrospective data-only study was approved by the KPSC institutional review board (#5815) and exempted from the need to obtain informed consent.

We developed 2 algorithms to determine whether an EHR approach using ICD-9 codes would reliably capture and predict biopsy-proven MN. The first

**Author Affiliations**

[1] Division of Nephrology and Hypertension, Los Angeles Medical Center, CA
[2] Department of Research and Evaluation, Kaiser Permanente Southern California, Pasadena, CA
[3] Department of Renal Pathology, Los Angeles Medical Center, CA
[4] Department of Nephrology, Stanford University, Palo Alto, CA

**Corresponding Author**

John J Sim, MD (john.j.sim@kp.org)

Keywords: electronic health records, epidemiology, membranous glomerulonephropathy, rare diseases

approach was the *clinically intuitive approach* (clinical approach), in which we used all ICD-9 codes that suggested a diagnosis of MN. These codes were 581.1 (nephrotic syndrome with lesion of membranous glomerulonephritis), 582.1 (chronic glomerulonephritis with lesion of membranous glomerulonephritis), and 583.1 (nephritis and nephropathy, not specified as acute or chronic, with lesion of membranous glomerulonephritis). MN was considered as "diagnosed" and predicted if at least 1 selected code appeared in the patient's EHR within the 1 year after kidney biopsy.

The second approach was the *data-driven/agnostic approach* (data-driven approach). In this method, we evaluated all ICD-9 codes for each patient who had biopsy-proven MN. We then determined which of the ICD-9 codes had the strongest associations with MN collected from supervised learning results using R software (version 3.4.3).[10] The learning was realized using multiple logistic modeling in a randomly resampled subset with half the size of the cohort, and a total of 1000 resampling sets were accomplished. Thus, for each time of sampling, half of the patients were randomly selected for training, and the other half were used for validation. All ICD-9 codes appearing in the EHR during the year after kidney biopsy were included in the modeling and tested for association with MN. Demographic and clinical variables (age, sex, race, birth year, biopsy year, and hospital) were included in the models as covariates to adjust for potential confounding effects.

We calculated the sensitivity, specificity, and positive predictive value (PPV) of the 2 approaches. Sensitivity analyses were performed, whereby we required at least 2 or at least 3 ICD-9 code appearances for case identification. Finally, to examine coding practices, we determined the cumulative incidence of ICD-9 codes for MN more than 12 months and compared those with biopsy-proven MN diagnoses vs those with alternate diagnoses.

## RESULTS

The study cohort comprised 416 patients who had biopsy-proven MN. Using the clinical approach, the sensitivity of at least 1 code for MN (581.1, 582.1, 583.1) in the first 12 months after kidney biopsy was 86% (78% if ≥ 2 codes and 70% if ≥ 3 codes). Specificity was 76% (83% and 87% for ≥ 2 and ≥ 3 codes, respectively), and PPV was 26% (31% and 35%, respectively; Table 1).

In the data-driven approach, a total of 6280 unique codes were found in the 1-year follow-up, and 127 unique codes remained for learning after excluding noninformative codes defined as prevalence less than 5%. We built 1000 different sets of subjects by randomly sampled half size of each biopsy-proven group, and a multiple logistic model was used to test associations in each resampled set. After collecting test results from all the resampled sets, we found that codes 581.1 and 583.1 were always at the top 10 strongest associations with biopsy-proven MN, with odds ratios of 2.62 (95% confidence interval = 1.77-3.86) and 3.77 (95% confidence interval = 2.54-5.6), respectively. The other code selected using the clinically intuitive approach (582.1) was found with a very low prevalence, at 1.3%, and a nonsignificant odds ratio

**Table 1. Clinical-based approach to identification of membranous nephropathy (MN); codes 581.1, 582.1, 583.1[a]**

| Codes reported ≥ 1 time | | | | Codes reported ≥ 2 times | | | | Codes reported ≥ 3 times | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Biopsy-proven MN | | | | Biopsy-proven MN | | | | Biopsy-proven MN | | | |
| Diagnosed MN | Yes | No | Total | Diagnosed MN | Yes | No | Total | Diagnosed MN | Yes | No | Total |
| Code reported ≥ 1 time | | | | | | | | | | | |
| Yes | 357 | 1023 | 1380 | Yes | 324 | 732 | 1056 | Yes | 293 | 554 | 847 |
| No | 59 | 3284 | 3343 | No | 92 | 3575 | 3667 | No | 123 | 3753 | 3876 |
| Total | 416 | 4307 | 4723 | Total | 416 | 4307 | 4723 | Total | 416 | 4307 | 4723 |
| Sensitivity (95% CI) | 0.86 (0.82-0.89) | | | Sensitivity (95% CI) | 0.78 (0.74-0.82) | | | Sensitivity (95% CI) | 0.70 (0.66-0.75) | | |
| Specificity (95% CI) | 0.76 (0.75-0.78) | | | Specificity (95% CI) | 0.83 (0.82-0.84) | | | Specificity (95% CI) | 0.87 (0.86-0.88) | | |
| PPV (95% CI) | 0.26 (0.24-0.28) | | | PPV (95% CI) | 0.31 (0.28-0.33) | | | PPV (95% CI) | 0.35 (0.31-0.38) | | |

[a] Code 581.1 is nephrotic syndrome with lesion of membranous glomerulonephritis; code 582.1 is chronic glomerulonephritis with lesion of membranous glomerulonephritis; and code 583.1 is nephritis and nephropathy, not specified as acute or chronic, with lesion of membranous glomerulonephritis.
CI = confidence interval; PPV = positive predictive value.

**Table 2. Data-driven approach to identification of membranous nephropathy (MN); codes 581.1, 583.1[a]**

| Codes reported ≥ 1 time | | | | Codes reported ≥ 2 times | | | | Codes reported ≥ 3 times | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Biopsy-proven MN | | | | Biopsy-proven MN | | | | Biopsy-proven MN | | | |
| Diagnosed MN | Yes | No | Total | Diagnosed MN | Yes | No | Total | Diagnosed MN | Yes | No | Total |
| Yes | 344 | 252 | 596 | Yes | 306 | 128 | 434 | Yes | 268 | 74 | 342 |
| No | 72 | 4055 | 4127 | No | 110 | 4179 | 4289 | No | 148 | 4233 | 4381 |
| Total | 416 | 4307 | 4723 | Total | 416 | 4307 | 4723 | Total | 416 | 4307 | 4723 |
| Sensitivity (95% CI) | 0.83 (0.79-0.86) | | | Sensitivity (95% CI) | 0.74 (0.69-0.78) | | | Sensitivity (95% CI) | 0.64 (0.60-0.69) | | |
| Specificity (95% CI) | 0.94 (0.93-0.95) | | | Specificity (95% CI) | 0.97 (0.97-0.98) | | | Specificity (95% CI) | 0.98 (0.98-0.99) | | |
| PPV (95% CI) | 0.58 (0.54-0.62) | | | PPV (95% CI) | 0.71 (0.66-0.75) | | | PPV (95% CI) | 0.78 (0.74-0.83) | | |

[a] See footnote to Table 1 for explanation of diagnosis codes.
CI = confidence interval; PPV = positive predictive value.

  

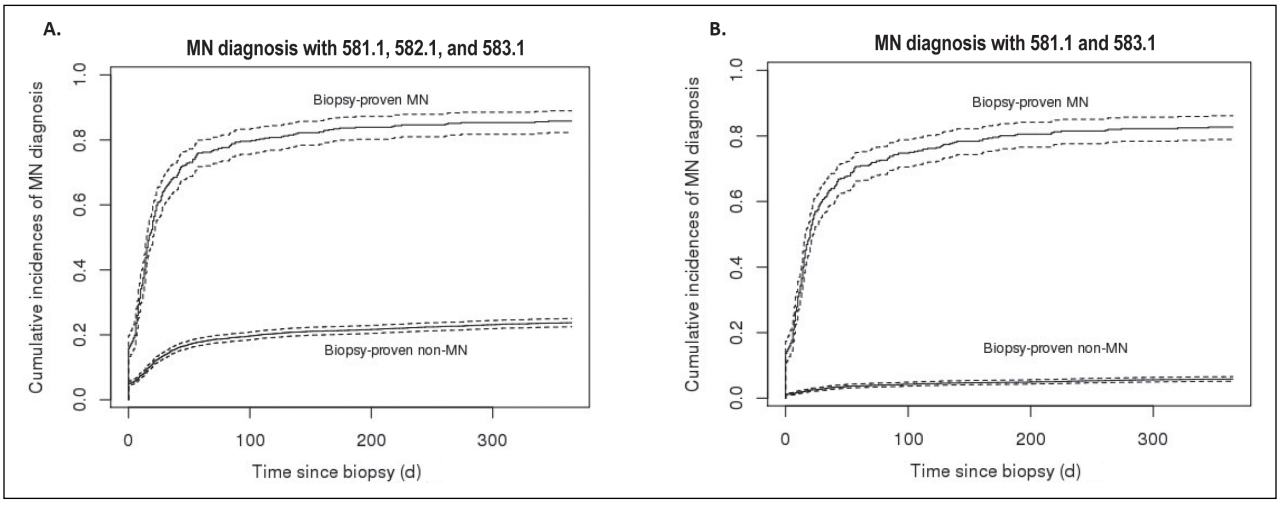


Figure 1. Cumulative incidences of coding for membranous nephropathy (MN) in the 365 days after kidney biopsy, comparing patients with biopsy-proven MN with those with biopsy-proven diagnoses other than MN, using either a clinically intuitive approach (A, on left) or a data-driven/agnostic approach (B, on right). Cumulative incidence curves with the clinical approach demonstrated that 64%, 81%, and 86% of patients with biopsy-proven MN received an International Classification of Diseases, Ninth Revision diagnosis code suggestive of MN by 30, 120, and 365 days. Cumulative incidences using data-driven/agnostic approach were marginally lower at 60%, 80%, and 83% of patients, respectively.

of 0.95 (95% confidence interval = 0.82-1.13). The third prevalent code found in the learning results was 583.81 but with low prevalence at 48.7%. Thus, we determined codes 581.1 and 583.1 as the 2 codes of interest for our data-driven approach, which happened to be 2 of the 3 ICD-9 codes suggesting MN used for the clinical approach. Using either of these 2 codes (ie, excluding code 582.1) appearing in the 12 months after biopsy led to substantial improvements in specificity (94%, 97%, and 98%, respectively, if ≥ 1, ≥ 2, or ≥ 3 ICD-9 codes were used) and PPV (58%, 71%, and 78%, respectively, if ≥ 1, ≥ 2, or ≥ 3 ICD-9 codes were used; Table 2).

Using the clinical approach, cumulative incidence curves demonstrated that among patients with biopsy-proven MN, 64%, 81%, and 86% received an ICD-9 code suggestive of MN by 30, 120, and 365 days, respectively, after kidney biopsy (Figure 1). Using the data-driven approach, cumulative incidences were marginally lower at 60%, 80%, and 83% of patients by 30, 120, and 365 days, respectively.

We performed further exploratory analyses of coding practices in non-MN cases. These results revealed that those with biopsy-proven focal segmental glomerulosclerosis accounted most for patients likely to receive an ICD-9 code suggesting MN (53% of false-positive cases).

## DISCUSSION

We evaluated the utility of diagnosis codes for identifying MN in a large integrated health system using an EHR-based approach. Our findings suggest that ICD-9 diagnosis codes might be a convenient tool to identify patients with MN using EHR and/or administrative claims information. Although the ICD-9-based clinical approach was reasonably sensitive and specific for MN (both > 80%), the relative rarity of MN cases even in this biopsied cohort resulted in a very poor predictive ability (PPV was only 26%, increasing to 35% if we required ≥ 3 code appearances). Were we to expand the cohort to all patients in this health system (in whom the incidence of MN approximates 10-20 per million population),[9] rather than restricting to those who had undergone a kidney biopsy, we would expect an even lower PPV. In contrast, using a data-driven approach, we could increase the PPV to 78%, albeit with a loss of sensitivity. Accordingly, each of the studied approaches has merit, depending on the goals of a research study: If estimating disease incidence, for example, an approach with high sensitivity (such that you are capturing most patients) should be selected. However, if studying disease natural history or treatment outcomes in patients with MN,[11] an approach with high PPV (such that you can be confident that you are studying cases of true MN) should be selected.

Our study has several limitations. First, we restricted our study population to people who had undergone a kidney biopsy: Accordingly, our outcome measures (sensitivity, specificity, and PPV) are applicable only to biopsied patients. Second, our cohort was composed of a racially/ethnically diverse cohort of patients with private insurance residing in Southern California.[9] Therefore, the applicability of our findings to other patient populations with potentially different disease distributions, and to other health systems with potentially different coding practices, cannot be determined. The clinical practice environment at KPSC has standardized and population-based practices for many chronic conditions, including chronic kidney disease, and has achieved high levels of success.[12-14] However, there are no internal diagnostic or treatment guidelines for glomerular diseases.

A third limitation is that we restricted our analyses to ICD-9 codes that were in use at the time of and up to 1 year after the date of the studied kidney biopsy procedures. Thus, we cannot speak for the sensitivity, specificity, or PPV of more contemporary ICD-10 codes. However, a major application of our

diagnostic algorithms (once validated in other populations) would be to study long-term patient outcomes (eg, disease natural history, treatment practices, and renal outcomes) among patients with MN. Therefore, identifying patients who underwent biopsy before October 2015 (when ICD-10 codes were implemented) using ICD-9 codes and then capturing outcomes using a combination of ICD-9 and ICD-10 codes would be most appropriate.[11]

Finally, the highest PPV we could achieve was 78%. Although it may be beyond the scope of this study (and potentially limiting clinical utility), additional measures to enhance PPV may include the addition of demographics, health care utilization data (eg, nephrology visits), medication data (eg, immunosuppressive therapies), laboratory data (eg, urine protein or phospholipase $A_2$ receptor antibody levels[15]), and/or free text data extracted from kidney biopsy reports.

## CONCLUSION

We found that ICD-9 diagnosis codes suggestive of MN received in the year after a kidney biopsy could be used to identify cases of MN with reasonable sensitivity but with lower PPV. A data-driven approach (whereby we were agnostic to the meaning of the ICD-9 diagnosis codes and instead selected those codes that were most strongly associated with a true diagnosis of MN) performed better than a clinical approach (where we focused only on those ICD-9 codes that suggested MN). Use of more than 1 code appearing in the patient record further improved PPV.

Our results suggest that ICD-9 diagnosis codes, if applied carefully or with support from machine learning, might be a convenient and reliable tool to identify large-scale populations of patients with MN using EHR or administrative/claims data. We welcome further studies aiming to validate and expand our approaches in other patient populations and to extend them to other glomerular diseases. Ultimately, applying these approaches to facilitate large-scale epidemiologic studies will inform understanding of the characteristics, experiences, and outcomes of real-world patients with MN. ❖

**Disclosure Statement**

*The author(s) have no conflicts of interest to disclose.*

**Acknowledgments**

*This study was funded and supported by the Kaiser Permanente Southern California (KPSC) Clinician Investigator Award (JJS) and by KPSC Regional Research.*

*Kathleen Louden, ELS, of Louden Health Communications performed a primary copy edit.*

**Authors' Contributions**

*John J Sim, MD, participated in the study concept and design; participated in data acquisition, analysis, and interpretation; contributed to drafting of the manuscript; and cosupervised the study. He had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Yu-Hsiang Shu, PhD, participated in the study concept and design; participated in data acquisition, analysis, and interpretation; and contributed to drafting of the manuscript. Amy Z Sun, MD, and Michelle M O'Shaughnessy, MD, participated in the study concept and design and contributed to drafting of the manuscript. Teresa N Harrison, SM, and Aviv Hever, MD, participated in data acquisition, analysis, and interpretation and helped critically revise the manuscript. Steven J Jacobsen, MD, PhD, helped critically revise the manuscript for important intellectual content; lent administrative, technical, or material support; and cosupervised the study.*

**How to Cite this Article**

Sun AZ, Shu Y-H, Harrison TN, et al. Identifying patients with rare disease using electronic health record data: The Kaiser Permanente Southern California Membranous Nephropathy Cohort. Perm J 2020;24:19.126. DOI: https://doi.org/10.7812/TPP/19.126

**References**

1. Nazareth TA, Kariburyo F, Kirkemo A, et al. Patients with idiopathic membranous nephropathy: A real-world clinical and economic analysis of US claims data. J Manag Care Spec Pharm 2019 Sep;25(9):1011-20. DOI: https://doi.org/10.18553/jmcp.2019.18456 PMID:31283419
2. Wetmore JB, Guo H, Liu J, Collins AJ, Gilbertson DT. The incidence, prevalence, and outcomes of glomerulonephritis derived from a large retrospective analysis. Kidney Int 2016 Oct;90(4):853-60. DOI: https://doi.org/10.1016/j.kint.2016.04.026 PMID:27425855
3. Moxey-Mims MM, Flessner MF, Holzman L, et al. Glomerular diseases: Registries and clinical trials. Clin J Am Soc Nephrol 2016 Dec 7;11(12):2234-43. DOI: https://doi.org/10.2215/CJN.00540116 PMID:27672219
4. O'Shaughnessy MM, Hogan SL, Poulton CJ, et al. Temporal and demographic trends in glomerular disease epidemiology in the southeastern United States, 1986-2015. Clin J Am Soc Nephrol 2017 Apr 3;12(4):614-23. DOI: https://doi.org/10.2215/CJN.10871016 PMID:28325866
5. O'Shaughnessy MM, Hogan SL, Thompson BD, Coppo R, Fogo AB, Jennette JC. Glomerular disease frequencies by race, sex and region: Results from the International Kidney Biopsy Survey. Nephrol Dial Transplant 2018 Apr 1;33(4):661-9. DOI: https://doi.org/10.1093/ndt/gfx189 PMID:29106637
6. Chibnik LB, Massarotti EM, Costenbader KH. Identification and validation of lupus nephritis cases using administrative data. Lupus 2010 May;19(6):741-3. DOI: https://doi.org/10.1177/0961203309356289 PMID:20179167
7. Himmelfarb J. Chronic kidney disease and the public health: Gaps in evidence from interventional trials. JAMA 2007 Jun 20;297(23):2630-3. DOI: https://doi.org/10.1001/jama.297.23.2630 PMID:17579231
8. O'Shaughnessy MM, Cheng XS, Montez-Rath ME, Lafayette RA, Winkelmayer WC. Validating identification of patients with small vessel vasculitis, with or without renal involvement, using administrative healthcare records. Clin Nephrol 2017 Mar;87(3):159-62. DOI: https://doi.org/10.5414/CN109035 PMID:28102817
9. Sim JJ, Batech M, Hever A, et al. Distribution of biopsy-proven presumed primary glomerulonephropathies in 2000-2011 among a racially and ethnically diverse us population. Am J Kidney Dis 2016 Oct;68(4):533-44. DOI: https://doi.org/10.1053/j.ajkd.2016.03.416 PMID:27138468
10. Team RCR. A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing [cited 2019 Dec 18]. Available from: www.R-project.org
11. Sim JJ, Bhandari SK, Batech M, et al. End-stage renal disease and mortality outcomes across different glomerulonephropathies in a large diverse US population. Mayo Clin Proc 2018 Feb;93(2):167-78. DOI: https://doi.org/10.1016/j.mayocp.2017.10.021 PMID:29395351
12. Sim JJ, Zhou H, Bhandari S, et al. Low systolic blood pressure from treatment and association with serious falls/syncope. Am J Prev Med 2018 Oct;55(4):488-96. DOI: https://doi.org/10.1016/j.amepre.2018.05.026 PMID:30166081
13. Danforth KN, Hahn EE, Slezak JM, et al. Follow-up of abnormal estimated GFR results within a large integrated health care delivery system: A mixed-methods study. Am J Kidney Dis 2019 Nov;74(5):589-600. DOI: https://doi.org/10.1053/j.ajkd.2019.05.003 PMID:31324445
14. Sim JJ, Rutkowski MP, Selevan DC, Batech M, Timmins R, Slezak JM, Jacobsen SJ, Kanter MH: Kaiser Permanente creatinine safety program: A mechanism to ensure widespread detection and care for chronic kidney disease. Am J Med 2015;128:1204-11.e1. DOI: https://doi.org/10.1016/j.amjmed.2015.05.037
15. Beck LH Jr, Bonegio RG, Lambeau G, et al. M-type phospholipase A2 receptor as target antigen in idiopathic membranous nephropathy. N Engl J Med 2009 Jul 2;361(1):11-21. DOI: https://doi.org/10.1056/NEJMoa0810457 PMID:19571279