Mini-Review

DOI: 10.5582/irdr.2020.03123

Different approaches to improve cohort identification using electronic health records: X-linked hypophosphatemia as an example

Jose Jesus Broseta

Department of Nephrology and Renal Transplantation, Hospital Clínic of Barcelona, Barcelona, Spain.

SUMMARY Electronic Health Records (EHRs) represent a source of high value data which is often underutilized because exploiting the information contained therein requires specialized techniques unavailable to the end user *i.e.* the physician or the investigator. Here I describe four simple and practical avenues that will allow the standard EHR end user to identify patient cohorts: the use of diagnostic codes from different international catalogues; a search in reports from complementary tests (*e.g.* radiographs or lab tests) for any result of interest; a free text search; or a drug prescription search in the patient's electronic prescription record. This medical approach is acquiring great importance in the field of rare diseases, and here I demonstrate its application with X-linked hypophosphatemia. The use of these four EHR questioning approaches makes finding a cohort of patients of any condition or disease feasible and manageable, and once each case record is checked, a well-defined cohort can be assembled.

Keywords Cohort identification, electronic health records, patient search, methodological approach, XLH

1. Introduction

Electronic Health Records (EHRs), of common use today, have changed the way clinical information is acquired and documented. They provide quick and economic access to vast amounts of data for clinical and research purposes. However, the exploitation of this information normally uses technical methods, such as natural language processing (NLP), text mining or machine learning techniques which in turn require specialized bioinformatic processes for their implementation (*1*). As a result, this vast amount of data may be inaccessible to the end users, such as clinicians or investigators, who are often looking for patients to recruit for clinical trials or their own observational studies (*2*). The absence of end-user tools greatly diminishes the EHRs' potential as a source of high value information into an underutilized resource that is exacerbated by the number of electronic health record software developers (*3*) and the lack of integration between them (*4*).

In recent years interest has especially grown in the area of rare diseases, which involves thousands of very different diseases that affect a small number of people within the general population (*5*). In Europe, rare diseases are defined as those with a prevalence of less than 1 in every 2,000 people, and within this group a disease is considered to be ultra-rare if it affects 1 in every 50,000 people (*6*). Despite the rising awareness of rare diseases, they still represent a challenge for clinicians, investigators, and public health systems (*7*). The small number of affected patients makes it challenging to investigate the pathophysiological mechanisms and discourages many companies from investing in treatments. In addition, few doctors recognize the heterogeneous signs and symptoms, leading to missed or delayed diagnoses. The length of time for the diagnosis of a rare disease is around 4.8 years, with the patient seeing an average of 7.3 different specialists for case evaluations (*8*). This delay has a serious impact on the patient's quality of life. Earlier identification of the disease is vital, not only for patient follow-up but also for furthering research.

Here, I describe different but compatible approaches that make it possible for a patient cohort to be identified by the standard EHR end user. Whether all or some of these approaches are workable will depend on the EHR developer and the information access provided by the software. For our example, I will apply the four different strategies to the rare disease, X-linked hypophosphatemia.

2. X-linked hypophosphatemia

The estimated incidence and prevalence of X-linked

hypophosphatemia (XLH) are less than 1 in 2,500 and 1 in 20,000, respectively (*9*). Despite the prevalence of this disorder being similar to other known pathologies, and the existence of specialists involved in its diagnosis and treatment, awareness and knowledge about this disease are still relatively low.

XLH has its origin in the mutation of the *PHEX* gene, with a dominant X-linked inherited trait. The mutated *PHEX* translates into an increased serum concentration of fibroblast growth factor 23 (FGF-23), a hormone whose main function is to regulate serum phosphate levels. The decreased clearance of FGF-23 leads to decreased tubular phosphate reabsorption which, together with decreased intestinal absorption of phosphorus, induces hypophosphatemia and ultimately the clinical manifestations of the disease (*10,11*).

The signs and symptoms, which can vary in severity, are evident from the first months of life. The most typical clinical features, which are soon noticeable in childhood, are bowed legs (*genu varum*) and short stature. But there may be other manifestations such as knock-knee (*genu valgum*), frequent fractures, osteoarticular pain, extraosseous calcifications, dental problems or hearing impairment, which in many cases cause a significant functional limitation that negatively impacts the patient's quality of life (*12*). Once suspicion of XLH is raised from the physical examination or the imaging tests, the diagnostic approach should begin by ruling out other causes of rickets, especially deficiencies. Patients with XLH have normal parathyroid hormone (PTH), vitamin D, and calcium in the blood and urine, but serum phosphate levels and tubular phosphate reabsorption are low (*11*).

Conventional treatment is based on oral phosphorus supplements and active vitamin D analogues to maintain alkaline phosphatase within normal range and minimize bone deformities, therefore early diagnosis is essential (*13*). However, these treatments are not without adverse effects, of which the most common are gastrointestinal disorders and the most serious are secondary hyperparathyroidism and nephrocalcinosis (*14*). The development of a new monoclonal antibody has raised new hopes in the management of the disease (*15,16*), as it has shown to stabilize serum phosphate levels and lead to improvements in rickets, skeletal healing, and physical function (*13*). Doctors need to take into account all their current and past patients who could benefit from any therapeutic breakthrough, and those who could participate in new clinical trials and observational studies. This is only possible if XLH patients can be correctly identified.

3. Methods for cohort identification

I demonstrate an EHR search of XLH patients with four easy-to-use data sources.

3.1. Diagnostic codes

One search method available in virtually all institutional systems is searching the diagnostic codes of international catalogues. However, a frequently encountered difficulty is that most catalogues group diseases together with other entities, so extensive case information is needed for correct screening, especially when dealing with rare diseases. This can be achieved using the other methods described below.

The first option is the widely used International Statistical Classification of Diseases and Related Health Problems (ICD) with its different editions: ICD-9, -10 and -11 are the editions most commonly used in hospitals. Many studies have been published using this classification for patient identification (*17-19*).

In ICD-9, XLH is considered a disorder of phosphorus metabolism and the name given is old terminology: vitamin D-resistant rickets, code 275.3. ICD-10 goes a step further in the classification of these disorders of phosphorus metabolism by a separation into four groups: XLH is here coded as E83.30, which encompasses unspecified disorders of phosphorus metabolism. However, ICD-11 made a major change in the classification. It removed XLH from the section of phosphorus metabolism disorders and placed it in the Disorders of vitamin D metabolism or transport, forming its own group, hypophosphatemic rickets, encoded as 5C63.22.

Another widely used classification tool is the Diagnosis-Related Group (DRG), which attempts to classify the diagnosis as reimbursable "products" provided by the hospital, which means that it is designed more for billing than for clinical data analysis (*20*). This system normally places rare diseases, such as XLH, in a general category, classified under "Inborn and other disorders of metabolism", coded as 642.

On the opposite end of the scale is the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) which is a coded, comprehensive, multilingual, clinical terminology that is becoming a standard in many countries (*21*). Here XLH is given a unique code: 82236004. Another classification database available is the Online Mendelian Inheritance in Man (OMIM), which catalogues all known human genetic disorders; XLH as a hereditary disease has its code: 307800. The summary of the codes for XLH is shown in Figure 1.

3.2. Complementary tests

Another way to find a cohort of patients among the big data provided by the EHRs is to look for diagnostic criteria already discussed in the lab test results or in the imaging reports. Once these findings have been filtered, each case must be reviewed to generate a differential diagnosis that will confirm or rule out the disease.

In the case of XLH, the most effective strategy would be to filter cases of binomial decrease in tubular phosphate reabsorption and hypophosphatemia. This

could raise suspicions of phosphopenic rickets. Even so, we would still have to rule out calcipenic rickets (also known as vitamin D-deficient rickets) by focusing on PTH and vitamin D levels, and Fanconi's syndrome, by looking at urine tests positive for glucosuria, high pH levels and elevated fractional excretion of potassium (*22*). Other approaches that follow this strategy would be to look for the typical radiographic lesions (metaphyseal cupping and flaring and physeal widening and irregularity of the long bones). Additionally, a search for *PHEX* gene mutations should be made in the genetic results database (Figure 2).

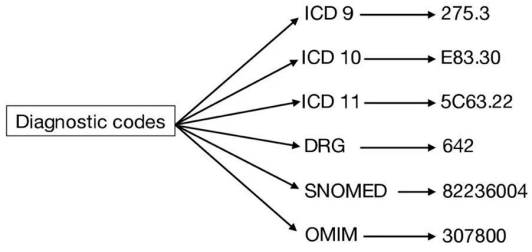

Figure 1. Diagnostic codes for X-linked hypophosphatemia. DRG, diagnosis-related group; ICD, International Statistical Classification of Diseases and Related Health Problems; OMIM, Online Mendelian Inheritance in Man; SNOMED, Systematized Nomenclature of Medicine.

3.3. Free text search

In most EHRs there is the possibility for the clinician to both code their findings in a structured format and also enter information in narrative free text. Unstructured free text searches, when available, can be more efficient than searching for a diagnostic code (*23*). There are different ways to search for free text. One way is to perform a plain text search by entering the different names given to the same disease. In the case of XLH, the different terms include: X-linked hypophosphatemic rickets, X-linked hypophosphatemia, vitamin D-resistant rickets, *etc.* Another alternative would be to search for text in clinical notes, imaging or genetic test reports. This strategy is more far-reaching but less specific than searching for diagnostic codes. It may be time-consuming for very prevalent diseases, but in the case of rare diseases it is best to begin with a long list of potential patients (*24*). Figure 3 shows an example of a free text search for X-linked hypophosphatemia.

An alternative or complementary approach is the use of natural language processing (NLP) to extract important information from text-based documents. This tool has proven to be more accurate in identifying

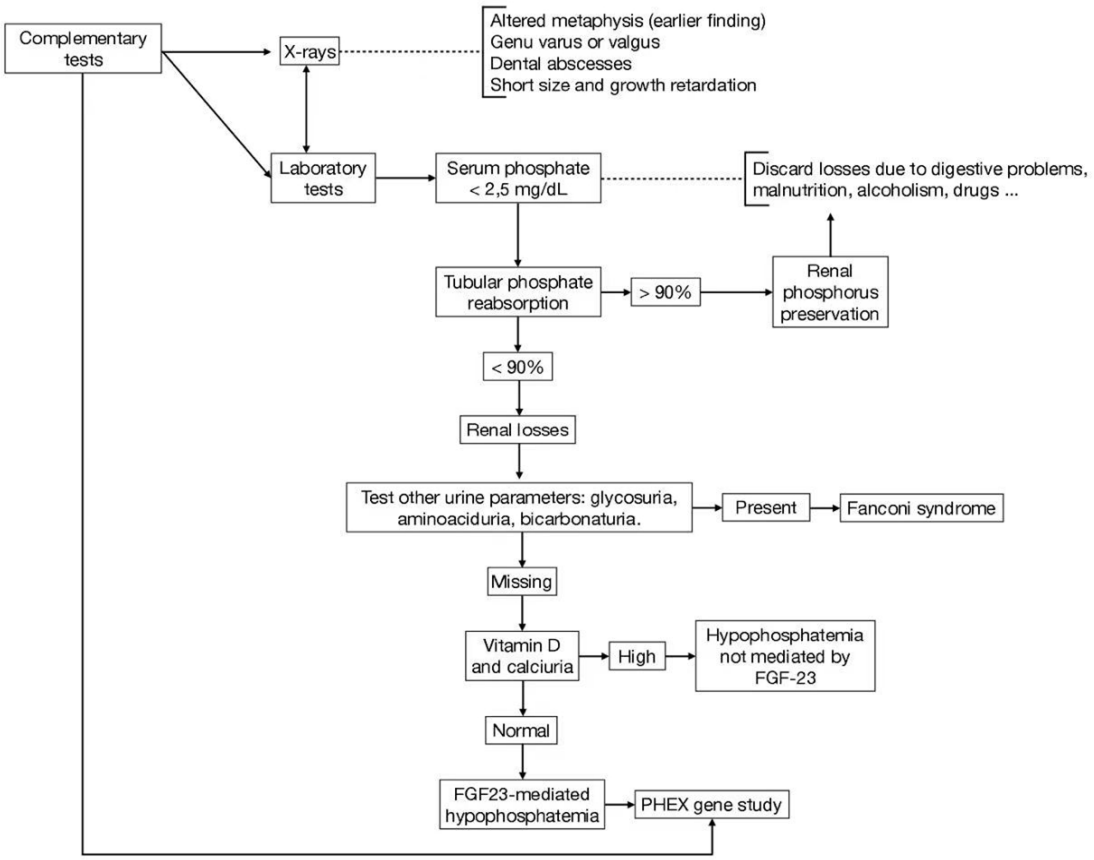

Figure 2. Complementary tests approach for X-linked hypophosphatemia.

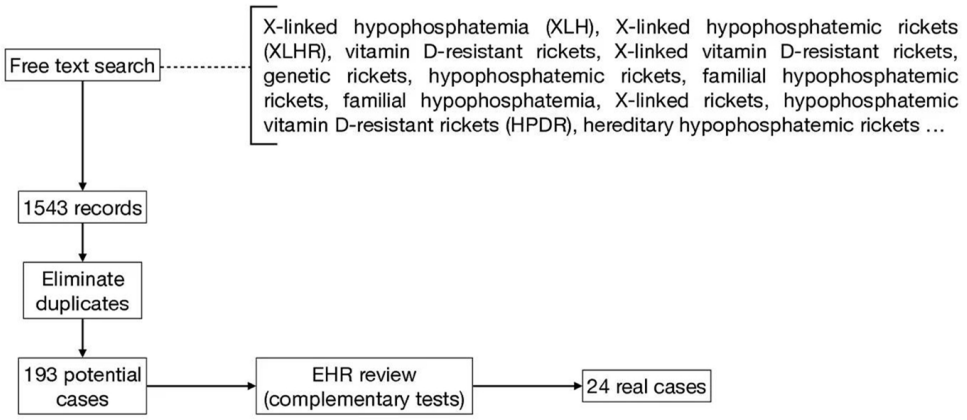

Figure 3. Free text search approach.

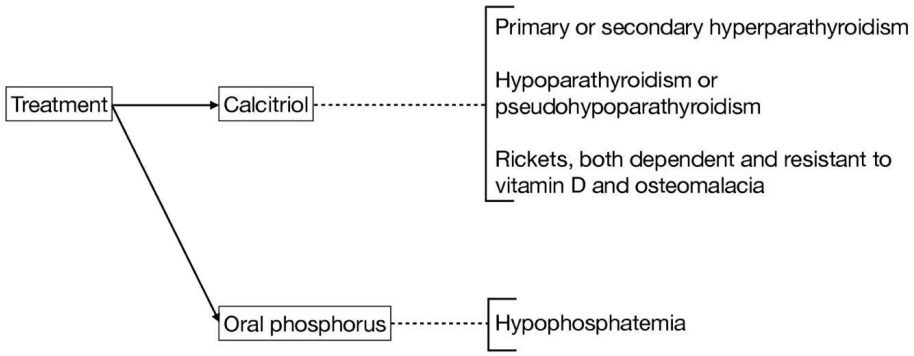

Figure 4. Patients' drug prescription approach.

postoperative complications compared with discharge coding information (*25*). It has also shown advantages in the classification of results of brain tumour magnetic resonance imaging reports in comparison to human expert classification, with excellent accuracy observed for tumour status classification (*26*). Unfortunately, free text searches are rarely exploited (*27,28*).

3.4. Electronic prescriptions

Finally, another way to find cohorts is to search for prescriptions in the patient's electronic medical records. Some drugs indicate specific pathologies, but if that is not the case for the disease being queried, the combination of drugs with several indications may sometimes lead to the disease in question. For example, this method is especially useful in XLH since the classical medical treatment is based on oral supplementation of phosphorus and vitamin D. These two medicinal products, often used separately, are prescribed together in very few situations (Figure 4). This approach can be a useful screening step to then be complemented by the review of each case to confirm the identification. It can be especially valid for those adult patients who received treatment in childhood and were subsequently lost to follow-up (*29*).

4. Conclusion

The identification of cohorts of patients with a certain condition or pathology using "EHR big data" is a growing concern (*1,30*). Sadly, this has become overly technical and out of the reach of doctors or investigators who are not experts in these new technologies (*31*). Therefore, for the EHR data interrogation to successfully satisfy the end user, easier ways are needed to obtain and analyse the information. I have introduced four different ways of finding patients that may be available to the entire medical community. I am aware that each institution with its own EHR software may not have all four of the proposed methods available. However, I am confident that just one or a combination of the methods

would facilitate the task of finding patients with a particular condition. In the case of a rare disease, such as XLH, it is preferable to rely on more sensitive methods, as an eventual "false positive" can always be discarded afterwards. With a well-defined cohort, any investigation or clinical trial can be launched (*32*).

Funding: This article has been commissioned and financially supported by Kyowa Kirin. Neither the author nor the founder have economic or patrimonial interests in the methodological approach explained in this article. The author states that the funder was not involved in the preparation of the manuscript. The author wishes to thank Anabel Herrero, PhD for providing editing assistance. Kyowa Kirin funded this assistance.

Conflict of Interest: JJB reports personal fees and non-financial support from Kyowa Kirin during the conduct of the study.

References

1. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, Lai AM. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inf Assoc. 2014; 21:221-230.
2. Schreiweis B, Trinczek B, Kopcke F, Leusch T, Majeed RW, Wenk J, Bergh B, Ohmann C, Rohrig R, Dugas M, Prokosch HU. Comparison of electronic health record system functionalities to support the patient recruitment process in clinical trials. Int J Med Inf. 2014; 83:860-868.
3. The Office of the National Coordinator for Health Information Technology. Health Care Professional Health IT Developers. 2017; *https://dashboard.healthit.gov/quickstats/pages/FIG-Vendors-of-EHRs-to-Participating-Professionals.php* (accessed August 11, 2020).
4. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012; 13:395-405.
5. Schieppati A, Henter JI, Daina E, Aperia A. Why rare diseases are an important medical and social issue. Lancet. 2008; 371:2039-2041.
6. Harari S. Why we should care about ultra-rare disease. Eur Respir Rev. 2016; 25:101-103.
7. Elliott E, Zurynski Y. Rare diseases are a "common" problem for clinicians. Aust Fam Physician. 2015; 44:630-633.
8. Engel P, Bagal S, Broback M, Boice N. Physician and patient perceptions regarding physician training in rare diseases: the need for stronger educational initiatives for physicians. J Rare Disord. 2013; 1:1-15.
9. Beck-Nielsen SS, Brock-Jacobsen B, Gram J, Brixen K, Jensen TK. Incidence and prevalence of nutritional and hereditary rickets in southern Denmark. Eur J Endocrinol. 2009; 160:491-497.
10. Gattineni J, Bates C, Twombley K, Dwarakanath V, Robinson ML, Goetz R, Mohammadi M, Baum M. FGF23 decreases renal NaPi-2a and NaPi-2c expression and induces hypophosphatemia *in vivo* predominantly *via* FGF receptor 1. Am J Physiol Ren Physiol. 2009; 297:F282-291.
11. Pavone V, Testa G, Gioitta Iachino S, Evola FR, Avondo S, Sessa G. Hypophosphatemic rickets: etiology, clinical features and treatment. Eur J Orthop Surg Traumatol. 2015; 25:221-226.
12. Chesher D, Oddy M, Darbar U, Sayal P, Casey A, Ryan A, Sechi A, Simister C, Waters A, Wedatilake Y, Lachmann RH, Murphy E. Outcome of adult patients with X-linked hypophosphatemia caused by *PHEX* gene mutations. J Inherit Metab Dis. 2018; 41:865-876.
13. Lambert AS, Zhukouskaya V, Rothenbuhler A, Linglart A. X-linked hypophosphatemia: Management and treatment prospects. Jt Bone Spine. 2019; 86:731-738.
14. Carpenter TO, Imel EA, Holm IA, Jan de Beur SM, Insogna KL. A clinician's guide to X-linked hypophosphatemia. J Bone Min Res. 2011; 26:1381-1388.
15. Carpenter TO, Whyte MP, Imel EA, Boot AM, Hogler W, Linglart A, Padidela R, Van't Hoff W, Mao M, Chen CY, Skrinar A, Kakkis E, San Martin J, Portale AA. Burosumab therapy in children with X-linked hypophosphatemia. N Engl J Med. 2018; 378:1987-1998.
16. Insogna KL, Briot K, Imel EA, *et al.* A randomized, double-Blind, placebo-controlled, phase 3 trial evaluating the efficacy of burosumab, an anti-FGF23 antibody, in adults with X-linked hypophosphatemia: week 24 primary analysis. J Bone Min Res. 2018; 33:1383-1393.
17. Boyd M, Specks U, Finkielman JD. Accuracy of the ICD-9 code for identification of patients with Wegener's granulomatosis. J Rheumatol. 2010; 37:474.
18. Smith JR, Jones FJS, Fureman BE, Buchhalter JR, Herman ST, Ayub N, McGraw C, Cash SS, Hoch DB, Moura LMVR. Accuracy of ICD-10-CM claims-based definitions for epilepsy and seizure type. Epilepsy Res. 2020; 166:106414.
19. Sun AZ, Shu Y-H, Harrison TN, Hever A, Jacobsen SJ, O'Shaughnessy MM, Sim JJ. Identifying patients with rare disease using electronic health record data: the Kaiser Permanente southern California membranous nephropathy cohort. Perm J. 2020; 24:19.126.
20. Tan JY-A, Senko C, Hughes B, Lwin Z, Bennett R, Power J, Thomson L. Weighted activity unit effect: evaluating the cost of diagnosis-related group coding. Intern Med J. 2020; 50:440-444.
21. Allones JL, Martinez D, Taboada M. Automated mapping of clinical terms into SNOMED-CT. An application to codify procedures in pathology. J Med Syst. 2014; 38:134.
22. González-Lamuño D. Hypophosphataemic rickets: diagnosis algorithm − how not to make a mistake. Adv Ther. 2020; 37:95-104.
23. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, Grannis SJ. Toward better public health reporting using existing off the shelf approaches: The value of medical dictionaries in automated cancer detection using plaintext medical data. J Biomed Inf. 2017; 69:160-176.
24. Maguire A, Johnson ME, Denning DW, Ferreira GLC, Cassidy A. Identifying rare diseases using electronic medical records: the example of allergic bronchopulmonary aspergillosis. Pharmacoepidemiol Drug Saf. 2017; 26:785-791.
25. Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K, Dittus RS, Rosen AK, Elkin PL, Brown SH, Speroff T. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA. 2011; 306:848-855.
26. Cheng LT, Zheng J, Savova GK, Erickson BJ. Discerning tumor status from unstructured MRI reports

− completeness of information in existing reports and utility of automated natural language processing. J Digit Imaging. 2010; 23:119-132.
27. Ford E, Nicholson A, Koeling R, Tate A, Carroll J, Axelrod L, Smith HE, Rait G, Davies KA, Petersen I, Williams T, Cassell JA. Optimising the use of electronic health records to estimate the incidence of rheumatoid arthritis in primary care: what information is hidden in free text? BMC Med Res Methodol. 2013; 13:105.
28. Price SJ, Stapley SA, Shephard E, Barraclough K, Hamilton WT. Is omission of free text records a possible source of data loss and bias in Clinical Practice Research Datalink studies? A case-control study. BMJ Open. 2016; 6:e011664-e011664.
29. Seefried L, Smyth M, Keen R, Harvengt P. Burden of disease associated with X-linked hypophosphataemia in adults: a systematic literature review. Osteoporos Int. 2021; 32:7-22.
30. Hemingway H, Asselbergs FW, Danesh J, Dobson R, Maniadakis N, Maggioni A, van Thiel GJM, Cronin M, Brobert G, Vardas P, Anker SD, Grobbee DE, Denaxas S. Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. Eur Hear J. 2018; 39:1481-1495.
31. Hruby GW, Matsoukas K, Cimino JJ, Weng C. Facilitating biomedical researchers' interrogation of electronic health record data: Ideas from outside of biomedical informatics. J Biomed Inf. 2016; 60:376-384.
32. Abrahão MTF, Nobre MRC, Gutierrez MA. A method for cohort selection of cardiovascular disease records from an electronic health record system. Int J Med Inform. 2017; 102:138-149.

Received September 28, 2020; Revised December 19, 2020; Accepted December 25, 2020.

Address correspondence to:
Jose Jesus Broseta, Department of Nephrology and Renal Transplantation, Hospital Clínic of Barcelona, Carrer de Villarroel 170, 08036 Barcelona, Spain.
E-mail: jjbroseta@clinic.cat

Released online in J-STAGE as advance publication January 17, 2021.