

Technical Report 2014-10-03

Estimation of prevalence of chronic kidney disease among diabetic patients in Austria

Milan Hronsky¹, Angelika Geroldinger¹, Georg Männer², Florian Endel³,
Gottfried Endel⁴, Rainer Oberbauer⁵, Georg Heinze¹

¹ Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Austria

² Department of Laboratory Medicine, Medical University of Vienna, Austria

³ Faculty of Mathematics and Geoinformatics, Technical University of Vienna, Austria

⁴ Main Association of Austrian Social Security Institutions, Vienna, Austria

⁵ Department of Medicine, Medical University of Vienna, Austria

ABSTRACT

Background. Health care claims data bases as run by health insurance providers contain rich information about patients' health status and morbidities. However, data on diagnoses is usually not directly available. This gap could be filled by linking health care claims databases with anonymous data on hospital admissions as provided by governmental organizations using probabilistic, i.e., de-personalized linkage methods. Extrapolation of the subset of hospitalized patients to the general population can be done by conditioning on the history of prescribed drugs, assuming that patients with equal patterns of prescribed drugs have similar outcome.

Methods. Using probabilistic linkage based on six descriptors, we linked data from hospital admission diagnoses to a health care data base provided by the Austrian Sickness Funds, evaluating the quality of linkage. Using logistic regression for high-dimensional predictor space, imposing a penalty on the quadratic norm of the regression coefficients, we modeled presence of a chronic kidney disease admission diagnosis as a function of the history of drug prescription preceding the hospital admission. Separate models were developed for different sex-age-groups. The resulting high-dimensional drug prescription models were then used to estimate the population-wide apparent prevalence of CKD. The apparent prevalences were finally calibrated by using laboratory data available on a subset of the patients which serve to define CKD stages.

Results. The positive predictive value of probabilistic linkage was 0,9786, indicating a high precision of our linkage algorithm. The relative frequency of a CKD diagnosis in hospital admission data was 6.33% (females, 6.31%, males, 6.36%). The extrapolated adjusted prevalence of CKD was 25.5% (females, 33.3%, males, 17.7%). Cross-validation revealed high predictive accuracy of our approach.

Conclusion. Using techniques of probabilistic linkage and high-dimensional modeling, we were able to estimate, with adequate accuracy, the prevalence of different, in particular early, stages of CKD among diabetic Austrians. Our method provides the potential for several extensions, such as the investigation of regional patterns of disease prevalence, or the development of drug prescription based prediction models for disease incidence.

CONTENTS

ABSTRACT.....	2
1 INTRODUCTION.....	4
2 METHODS.....	5
2.1 Data bases.....	5
2.2 Probabilistic linkage.....	5
2.3 Prevalence estimation.....	5
2.4 High-dimensional modeling.....	6
2.5 Apparent prevalence estimation.....	7
2.6 Prevalence adjusted for sensitivity and specificity.....	7
2.7 Comparing prevalences of different CKD stages.....	8
2.8 Regional prevalences.....	8
2.9 Software used.....	8
3 RESULTS.....	10
3.1 Description of diabetic patients.....	10
3.2 Linkage success.....	11
3.3 Modeling results.....	11
3.4 Model performance.....	12
3.5 Prevalence estimation in patients with laboratory data.....	13
3.6 Adjusted prevalence estimation.....	14
3.7 Regional differences in prevalence (Austria map).....	15
4 CONCLUDING REMARKS.....	16
5 REFERENCES.....	17

1 INTRODUCTION

Chronic kidney disease is a major public health problem affecting more than 50 million people worldwide. The Kidney Disease Outcomes and Quality Initiative (2002) defines chronic kidney disease as the presence of a marker of kidney damage such as proteinuria or a decreased glomerular filtration rate (GFR) for at least three months. Disease staging is based on the GFR, which, in clinical practice, is approximated using the serum creatinine level and some demographic parameters.

One of the leading causes of CKD in developed countries is Diabetes. This report focuses on the estimation of the prevalence of CKD in the Austrian diabetic population by the linkage of several large data bases.

We had access to the drug prescription data of almost all Austrians in the years 2006 and 2007. We restricted our data base to the diabetic population by identifying the subjects who were prescribed hypoglycaemic medicines. These data were exactly linked to all serum creatinine level measurements which were taken during the same years in the Vienna General Hospital main laboratory (3,790 linked records). Serum creatinine levels, if transformed to glomerular filtration rate estimates (eGFR), allow to determine presence and stage of kidney disease. However, single measurements of eGFR cannot discriminate chronic kidney disease from acute kidney failure. Therefore, we used data on hospital discharge diagnoses which were probabilistically linked to the prescription data base (76,564 linked records), which allowed us to derive CKD and non-chronic kidney disease probabilities conditional on a patient's drug prescription pattern. These probabilities were then calibrated to indicate stages of chronic kidney disease using the data from the laboratory. At this point, we had achieved prevalence estimates of CKD stages but based on the diabetic subjects with laboratory measurements, who might not be representative for the whole Austrian diabetic population. This selection was finally corrected by some kind of inverse probability weighting. All analyses were performed separately for sex/age groups defined by bins of 15 years.

The remainder of the report is organized as follows: a methods section describes the data bases used and the steps involved in prevalence estimation. Detailed results of the estimation are given in the subsequent section. Here, the total and age/sex specific prevalence estimates are given along with resampling-based standard errors, and the high-dimensional model which estimates chronic kidney disease conditional on drug prescription patterns is analyzed in detail, showing relevance of various prescribed substances for predicting the CKD diagnosis status of a patient. Finally, regional differences in CKD prevalence in Austria are displayed in a spatial map.

2 METHODS

2.1 Data bases

The prescription data base (PDB) of the Main Association of Austrian Social Insurance Institutions (MAASII) holds data on drug prescription in general practitioners for all patients with health insurance in Austria. The data base covers more than 90% of the insured population in Austria. Available data consist of encrypted person identifier, encrypted identifier of the health insurance carrier, type of drug (ATC), quantity, and ingoing date of prescription. Two further tables describe the insured population by sex, birth year and residential district, and hold basic data on hospital admissions (date of discharge, length of stay, main diagnosis at discharge). The minimum basic data set (MBDS) provided by the Austrian Ministry of Health holds data on hospital admissions, with main and associated diagnoses, date of discharge, length of stay, and demographic descriptors (sex, birth year, residential district). Furthermore, a third data base was constructed by extracting all laboratory data from the diabetes patients that were available in the Vienna General Hospital main laboratory. Laboratory data were exactly linked to the PDB using the encrypted person identifiers. Data from 76 564 patients could be linked.

2.2 Probabilistic linkage

PDB and MBDS were linked using probabilistic linkage by making use of the following attributes which were available in both PDB and MBDS: sex, birth year, residential district, date of discharge, length of stay, main diagnosis at discharge.

We estimated the positive predictive value, i.e. the proportion of hospital stays that were correctly linked to records in the PDB relative to the total number of linked hospital stays, using the Duplicate Method described by Blakely and Salmond (2002).

2.3 Prevalence estimation

First, the population of diabetic patients in Austria was extracted from the PDB by defining all patients as diabetic who received prescriptions on drugs used in diabetes (ATC codes starting with A10). A proportion of this diabetic population is already in the end stage of renal disease and on renal replacement therapy, either by dialysis or kidney transplantation. These patients are known as they are registered in the Austrian Dialysis and Transplant Register. Pseudonymized identifiers of these patients were generated from the social insurance numbers by MAASII and then those patients were identified in the diabetic population. For this subgroup of patients, the estimated prevalence of CKD is assumed 1. For the remaining group of patients (the ‘prevalence population’, PP), we estimated the prevalence of CKD using high-dimensional modeling as described below.

2.4 High-dimensional modeling

A subgroup of the PP had been admitted to hospital during the evaluated time period. For these, diagnosis of CKD and of acute kidney disease (AKD) is available at discharge. We defined the CKD statuses of the persons admitted to hospital as ‚present‘ if in any hospital discharge within the evaluated time period, a CKD diagnosis (ICD10 code N18) was found, and as ‚absent‘ if no such diagnosis was found. The AKD status was defined as ‚present‘ if the CKD status was absent and the diagnoses contained the ICD10 codes N17 or N19, and as ‚absent‘ else. If several periods of hospital stays were found, we selected one as the index stay. Now we extracted all prescriptions of the hospitalized patients obtained in a time period of 3-6 months before the index hospital admission. Carriers in Austria have different billing cycles, ranging from day-accurate billing to accounting in three-months periods. Prescriptions that were filled after the hospital admission date but for which it was unclear whether they were prescribed before or after hospital admission, were not considered in this model. Both outcomes, CKD and AKD status, were modeled using logistic regression with all available ATCs as binary variables (prescribed or not prescribed). The analysis of the CKD status was based on all patients in the PP who had been admitted to hospital, whereas the analysis of the AKD status was only based on the subgroup of hospitalized patients in the PP with absent CKD status. For patients who were not hospitalized in the evaluated time period, we can estimate a probability of a present CKD status (P_{CKD}) and a probability of an AKD diagnosis assuming the patient has an absent CKD status ($P_{AKD|CKD}$) by applying the two logistic regression models to their prescribed ATC pattern in a randomly selected time period of 6 months.

For predicting the CKD and AKD status in hospital diagnoses we made use of the logistic ridge regression model (Le Cessie and van Houwelingen, 1992). This model accounts for possible overfit by imposing a penalty on the log likelihood which is equal to the sum of squared standardized regression coefficients multiplied by a tuning parameter. The tuning parameter was optimized by maximizing the ten-fold cross-validated log likelihood. Models were developed separately for sex/age groups defined by bins of 15 years. Ridge regression models have the advantage over conventional regression models that they can deal with both high collinearity and high dimensionality of predictors without running the risk to overfit the model on the data at hand, which would result in poor generalizability. Calibration (agreement of predicted and observed probability) and discrimination (by means of the concordance index) was evaluated by cross-validation, selecting nine tenths of the data as training set on which the model was developed and one tenth of the data as test set, repeating this process such that each observational unit has appeared in the test set once.

In a secondary analysis we modeled the CKD and AKD status using LASSO-type penalized logistic regression (Tibshirani, 1996). For LASSO, the penalty is defined as the sum of the absolute values of the standardized regression coefficients. While there is some advantage of the ridge-type penalty over the LASSO for prognostic models (Ambler et al., 2012), the LASSO yields sparse solutions, i.e. some regression coefficients are estimated as exactly zero. Thus, in order to calculate the probabilities P_{CKD} and $P_{AKD|CKD}$ predicted by the LASSO for a given patient, one only has to know the prescriptions of the patients for a reduced set of ATCs.

Prescriptions were only considered on a binary (present/absent) basis. Quantitative descriptors (number of prescribed defined daily doses) did not relevantly improve our models.

2.5 Apparent prevalence estimation

The fitted models were then applied to the population for whom no data on hospital admissions was available to calculate P_{CKD} and $P_{AKD|CKD}$ based on the prescribed drugs. For this analysis, we selected an index date randomly and computed the predicted probabilities using the available information on drugs prescribed before the index date. The predicted probabilities computed in the non-hospitalized persons and the true statuses in hospitalized persons (e.g., 1 for a CKD discharge diagnosis, 0 for no such diagnosis in the case of CKD) were averaged over the complete insured population to yield sex- and age group-specific prevalence rates. The predicted probability for kidney disease P_{KD} was then calculated as $P_{KD} = P_{CKD} + (1 - P_{CKD}) P_{AKD|CKD}$.

Variability of the prevalence estimates in sex/age groups was assessed by computing bootstrap standard errors which were obtained by repeating the prevalence estimation, including re-developing the logistic ridge regression models, on 50 resampled data sets.

2.6 Prevalence adjusted for sensitivity and specificity

From the laboratory data, we obtained information on serum creatinine and albuminuria, and computed the CKD stage based on MDRD formula for estimated glomerular filtration rate (eGFR) for each patient, defining stage 1 as eGFR > 90 and presence of albuminuria (≥ 30 mg/g), stage 2 as eGFR between 60 and 89, and presence of albuminuria (≥ 30 mg/g), stage 3a as eGFR between 45 and 59, stage 3b as eGFR between 30 and 44, stage 4 as eGFR between 15 and 29, and stage 5 as eGFR below 15. These data were used to calibrate the apparent prevalence estimation for CKD stage, computing adjusted cumulative prevalence (CP) rates which correct the apparent prevalences P_{KD} using the formula $CP_{adj,x} = \frac{P_{KD} - (1 - Sp_x)}{Se_x - (1 - Sp_x)} \cdot \frac{P_{CKD}}{P_{KD}}$, where Sp_x and Se_x are the specificity and the sensitivity of our apparent prevalence for discriminating patients with CKD stage $\geq x$ from patients with CKD stages $< x$, $x=1, 2, 3a, 3b, 4, 5$. From the group of patients for which laboratory data were available, Sp_x and Se_x were computed as follows: let TP_x and FN_x denote the sum of apparent prevalences P_{KD} and the sum of $(1 - P_{KD})$ over all patients with CKD stages $\geq x$, respectively. Se_x was defined as $TP_x / (TP_x + FN_x)$. Likewise, let FP_x and TN_x denote the sum of P_{KD} and the sum of $(1 - P_{KD})$ over all patients with CKD stages $< x$, respectively. Sp_x was defined as $TN_x / (TN_x + FP_x)$. Adjusted cumulative prevalences $CP_{adj,x}$ were computed for each CKD stage x and for each age/sex group. The correction factor P_{CKD} / P_{KD} was used to discount the proportion of patients with absent CKD status but present AKD status. Adjusted prevalences $P_{adj,x}$ for each CKD stage x were obtained by $P_{adj,x} = CP_{adj,x-1} - CP_{adj,x}$. Since the patients with laboratory data may not be representative for the full diabetic population, we reweighted these patients by assigning weights according to their inverse sampling probabilities (ISP). The ISP were obtained by first dividing the distribution of observed values of P_{KD} in the laboratory patient group into deciles, and then computing the proportions of patients of the prevalence population with P_{KD} falling into these

deciles, denoted by $D_1, \dots, D_{10}$. (If laboratory patients were representative for the prevalence population with respect to their apparent prevalence, these numbers were all around 0.1.) The weights for the laboratory patients in deciles 1, ..., 10 were defined as $D_1/0.1, \dots, D_{10}/0.1$, respectively. Reweighting was used to compute the values of Sp_x and Se_x .

Standard errors for $P_{adj,x}$ were obtained by bootstrap resampling of 50 samples of the original data set used to estimate the logistic ridge regression model with replacement, mapped to 50 resamples of the laboratory data set with replacement, and repeating the computation of Sp_x , Se_x , $CP_{adj,x}$ and $P_{adj,x}$ for each of the resamples. The empirical standard deviation over the 50 resampled versions of $P_{adj,x}$ serves as estimated standard error.

All these computations were done for each sex/15-years age group.

2.7 Comparing prevalences of different CKD stages

In order to estimate 95% confidence intervals (95% CI) for the ratios $P_{adj,x}/P_{adj,y}$ ($x, y = 3a, 3b, 4$ or 5) we first computed the effective sample sizes $N_{eff,x}$ and $N_{eff,y}$, i.e. $N_{eff,*} = P_{adj,*} (1 - P_{adj,*}) / SE_{emp,*}^2$ with $SE_{emp,*}$ the empirical standard error of $P_{adj,*}$ estimated by resampling. Next, the effective numerators $Z_{eff,x}$ and $Z_{eff,y}$ were determined as the numbers satisfying $P_{adj,*} = Z_{eff,*} / N_{eff,*}$. At this point, we made use of the following fact: If X and Y are independently binomially distributed variables based on sample sizes m and n and parameters p_1 and p_2 , respectively, then the random variable $\log(\frac{X}{m} / \frac{Y}{n})$ is approximately normally distributed and its variance can be estimated as $x^{-1} - m^{-1} + y^{-1} - n^{-1}$. Thus, we find that the variance $V_{x,y}$ of the logarithm of the ratio $P_{adj,x} / P_{adj,y}$ can be estimated as $Z_{eff,x}^{-1} - N_{eff,x}^{-1} + Z_{eff,y}^{-1} - N_{eff,y}^{-1}$. The upper and lower bounds of a 95% confidence interval of $\log(P_{adj,x}/P_{adj,y})$ are then given by $\log(P_{adj,x}/P_{adj,y}) \pm 1.96 \sqrt{V_{x,y}}$, respectively. Taking the exponential finally yields the 95% CI $(P_{adj,x}/P_{adj,y} e^{-1.96 \sqrt{V_{x,y}}}, P_{adj,x}/P_{adj,y} e^{1.96 \sqrt{V_{x,y}}})$ for the ratio of prevalences $P_{adj,x}/P_{adj,y}$.

2.8 Regional prevalences

Applying the coefficients of the logistic ridge regression models predicting P_{CKD} and $P_{AKD|CKD}$ to the ATC prescriptions patterns of the patients in a certain district, we can estimate district specific apparent prevalences for each sex/age group. Again, these prevalences can be adjusted for sensitivity and specificity using adequately weighted laboratory data. These regional prevalences were then standardized using indirect standardization (Inskip, 2000) by a reference population defined by sex and age group frequencies as observed in the total Austrian diabetic population.

2.9 Software used

Data base operations were performed using PostgreSQL. Data were extracted first to SAS (Version 9.3, 2011 SAS Institute Inc., Cary, NC, USA), in which the data were further pre-processed. Logistic ridge regression modeling was done in R using the glmnet package (Friedman et al., 2010).

Estimated coefficients were imported into PostgreSQL, in which the predicted probabilities of CKD were computed for all persons. The prevalence maps were again computed using R (Version 2.12.2, www.r-project.org).

3 RESULTS

3.1 Description of diabetic patients

Fig. 1A shows the age and sex distributions in the diabetic population (N=319,548). Panels 1B and 1C show the respective plots for the subgroups of patients with available laboratory data (N=3,790) and the patients listed in the dialysis and transplant registry (N=1,791).

Table 1 contains the numbers of patients and of explanatory variables (ATCs) used in the logistic regression models for the different age/sex groups.

Fig. 1: Age and sex distribution in the diabetic population and the two subpopulations

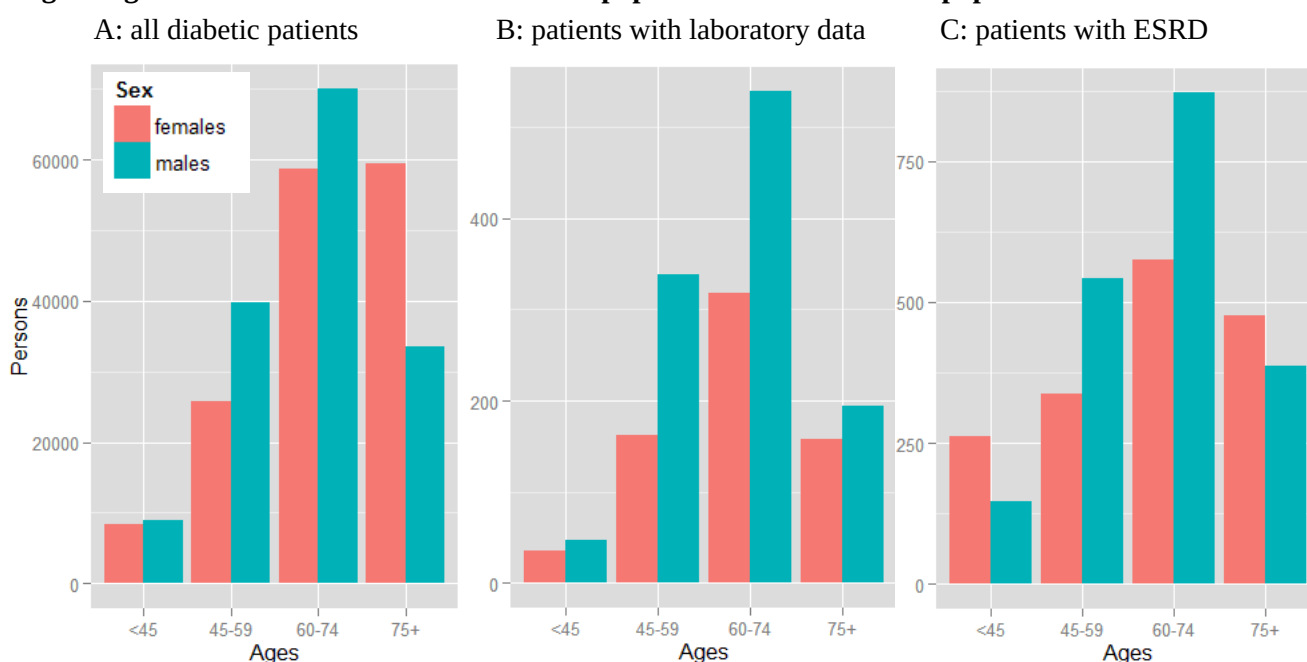


Table 1: Raw frequencies and number of covariables (ATCs) used in the logistic regression model, by age group and sex

Age	Sex	N	CKD	ATCs
<45	Male	2,624	32	165
	Female	3,328	11	223
45-59	Male	10,686	239	418
	Female	7,187	115	448
60-74	Male	21,751	1,318	529
	Female	18,211	786	547
>75	Male	13,764	1,514	465
	Female	25,230	2,493	580

3.2 Linkage success

The positive predictive value of record linkage was estimated at 0,9786.

3.3 Modeling results

Fig. 2 shows the magnitude of the regression coefficients in the CKD-models for each ATC, exemplarily for the age group 60-74 years. Solid lines connect the coefficients of the ridge models and circles correspond to the selected ATCs from the LASSO models. Fig. 3 lists the five ATCs with the greatest ridge regression coefficients and the five ATCs with the smallest ridge regression coefficients for both sexes and the age group 60-74 years. Short descriptions of these ATCs can be found in the list below the figure. Table 2 contains the apparent CKD-prevalences estimated in the ridge regression models for each age/sex group.

Fig. 2: Regression coefficients of the ridge (solid line) and the LASSO (circles) models for CKD

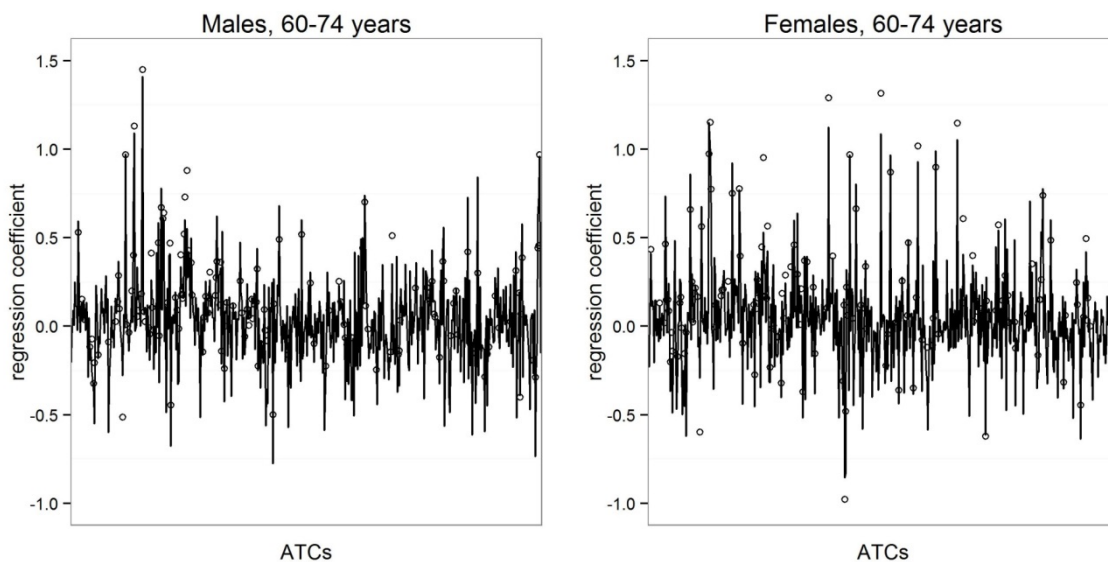
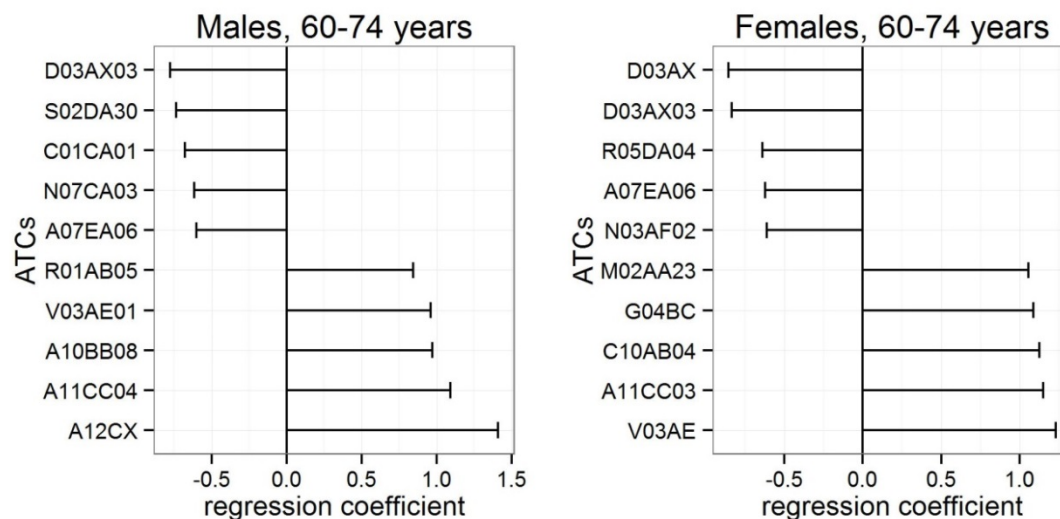


Fig. 3: Most relevant ATC-codes according to the logistic ridge regression model for CKD



Descriptions of ATCs listed above (↓ and ↑ indicate association of CKD with lower and higher prevalence, respectively)

MALES, 60-74 YEARS:

- ↓ D03AX03 dexpantenol (*preparations for treatment of wounds and ulcers*)
- ↓ S02DA30 analgesics and anesthetics, combinations (*otologicals*)
- ↓ C01CA01 etilefrine (*cardiac therapy*)
- ↓ N07CA03 flunarizine (*nervous system drugs*)
- ↓ A07EA06 budesonide (*antidiarrheals, intestinal antiinflammatory/antiinfective agents*)
- ↑ R01AB05 ephedrine (*nasal preparations*)
- ↑ V03AE01 polystyrene sulfonate (*drugs for treatment of hyperkalemia and hyperphosphatemia*)
- ↑ A10BB08 gliquidone (*drugs used in diabetes*)
- ↑ A11CC04 calcitriol (*vitamins*)
- ↑ A12CX mineral products different from Sodium, Zinc, Magnesium, Fluoride and Selenium (*mineral supplements*)

FEMALES, 60-74 YEARS:

- ↓ D03AX cicatrizants different from cod-liver oil ointments (*preparations for treatment of wounds and ulcers*)
- ↓ D03AX03 dexpantenol (*preparations for treatment of wounds and ulcers*)
- ↓ R05DA04 codeine (*cough and cold preparations*)
- ↓ A07EA06 budesonide (*antidiarrheals, intestinal antiinflammatory/antiinfective agents*)
- ↓ N03AF02 oxcarbazepine (*antiepileptics*)
- ↑ M02AA23 indometacin (*topical products for joint and muscular pain*)
- ↑ G04BC urinary concrement solvents (*urologicals*)
- ↑ C10AB04 gemfibrozil (*lipid modifying agents*)
- ↑ A11CC03 alfacalcidol (*vitamins*)
- ↑ V03AE drugs for treatment of hyperkalemia and hyperphosphatemia

Table 2: Apparent prevalences for CKD calculated from the ridge regression model

Age	Male	Female
<45	0.9%	0.3%
45-59	1.5%	1.0%
60-74	4.1%	2.9%
>75	8.9%	7.8%

3.4 Model performance

For simplicity, the description of the model performances is restricted to the ridge regression modeling the CKD status. Table 3 contains the c-indices for each age/sex group. Fig. 4 compares the predicted CKD-probabilities between the hospitalized patients with absent and the ones with present CKD status, exemplarily for the patients aged between 60 and 74 years. Ten-fold cross-validated calibration curves for the age group 60-74 years can be found in Fig. 5.

Table 3: c-indices for logistic ridge regression for CKD

Age	Male	Female
<45	0.949	0.988
45-59	0.914	0.942
60-74	0.841	0.86
>75	0.784	0.775

Fig. 4: Predicted CKD-probabilities of the hospitalized patients

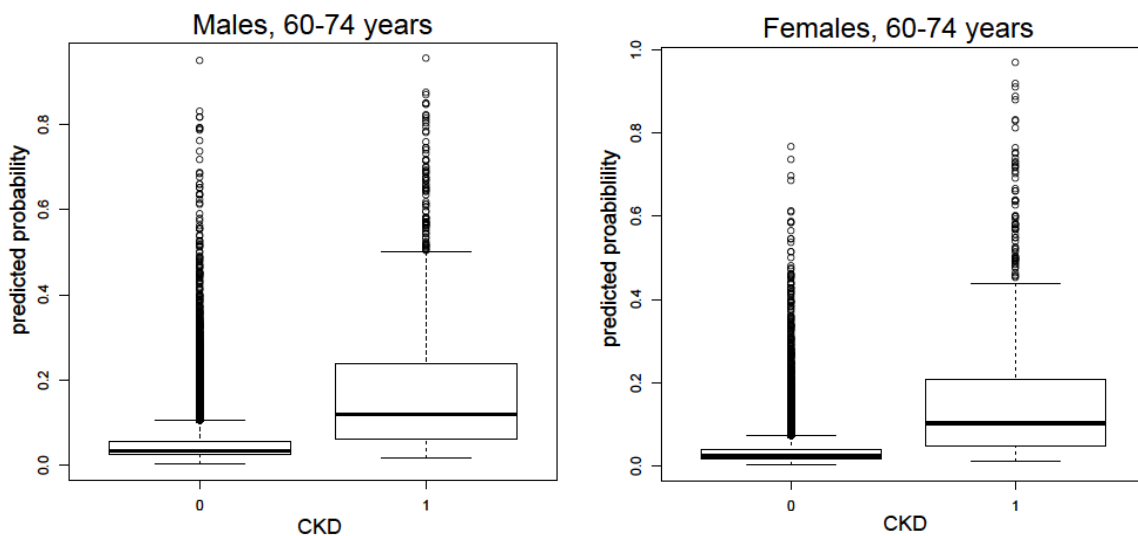
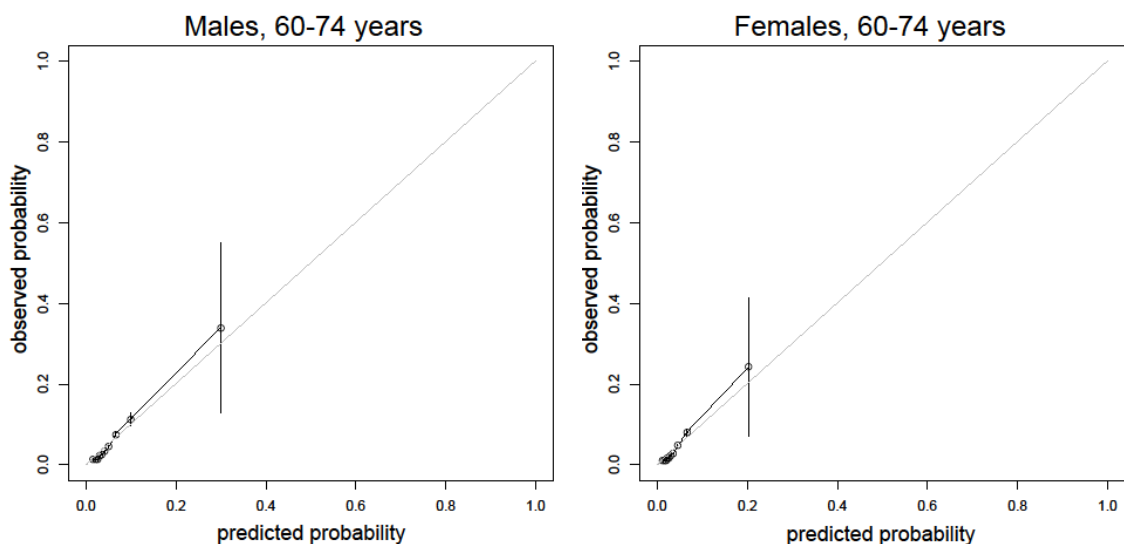


Fig. 5: Cross-validated calibration plots for the ridge regressions modeling CKD

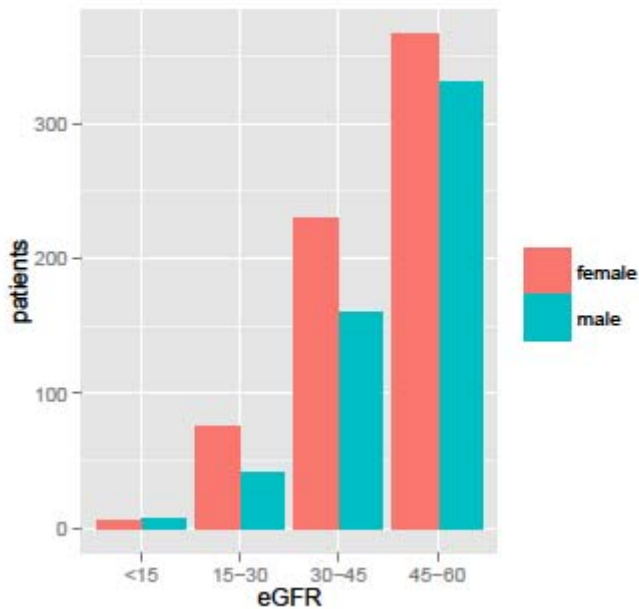


3.5 Prevalence estimation in patients with laboratory data

In the patients with available laboratory data, we observed a quite low prevalence of patients with micro- or macroalbuminuria. Thus, the prevalence computation was restricted to CKD of stage 3 or

higher. Fig. 6 shows the distribution of patients in the laboratory population by sex and stage of CKD for the stages 3a, 3b, 4 and 5.

Fig. 6: Numbers of patients with laboratory data by stage of CKD restricted to stage 3a, 3b, 4 and 5.



3.6 Adjusted prevalence estimation

The overall adjusted prevalence of CKD stages, based on laboratory data and patients enregistered in the dialysis and transplant registry, was estimated as 25.5% (standard error, 0.7%). The sex-specific prevalences were 17.7% (SE, 0.9%) for men and 33.3% (SE, 1.1%) for women. The stage-specific prevalences are presented in Tables 4 and 5. In particular, the prevalence for stage 3 was found to be 30.81 times (95% CI: 25.43, 37.32) higher than the prevalence for ESRD. The prevalence for stage 4 was estimated to be 2.66 times (95% CI: 2.03, 3.49) higher than the one for ESRD.

Table 4: Estimated stage specific prevalences of CKD among diabetic patients

Sex	stage 3a		stage 3b		stage 4		ESRD (stage 5)		CKD	
	preval	SE	preval	SE	preval	SE	preval	SE	preval	SE
All patients	14.9%	0.7%	7.9%	0.4%	1.97%	0.2%	0.74%	0.07%	25.5%	0.7%
male	11.2%	0.8%	4.6%	0.5%	1.06%	0.2%	0.87%	0.09%	17.7%	0.9%
female	18.5%	1.1%	11.3%	0.8%	2.87%	0.3%	0.61%	0.10%	33.3%	1.1%

Table 5: Estimated stage specific prevalences of CKD among diabetic patients for the different age groups

Age group	Sex	stage 3a		stage 3b		stage 4		ESRD (stage 5)	
		preval	SE	preval	SE	preval	SE	preval	SE
<45	male	1.8%	0.9%	0.5%	0.5%	0.2%	0.2%	0.4%	<0.1%
	female	0.7%	0.6%	0.9%	0.6%	0.3%	0.3%	0.3%	<0.1%
45-59	male	4.5%	0.8%	1.1%	0.5%	0.4%	0.2%	0.9%	0.1%
	female	8.3%	2.2%	2.1	0.7%	1.2%	0.6%	0.9%	0.3%
60-74	male	11.8%	1.1%	4.7%	0.7%	1.1%	0.3%	0.9%	0.1%
	female	20.0%	1.9%	7.4%	1.1%	1.2%	0.3%	0.7%	0.1%
>75	male	21.2%	1.9%	9.8%	1.5%	2.2%	0.7%	0.9%	0.3%
	female	24.5%	1.9%	20.9%	1.9%	5.7%	0.9%	0.4%	0.2%

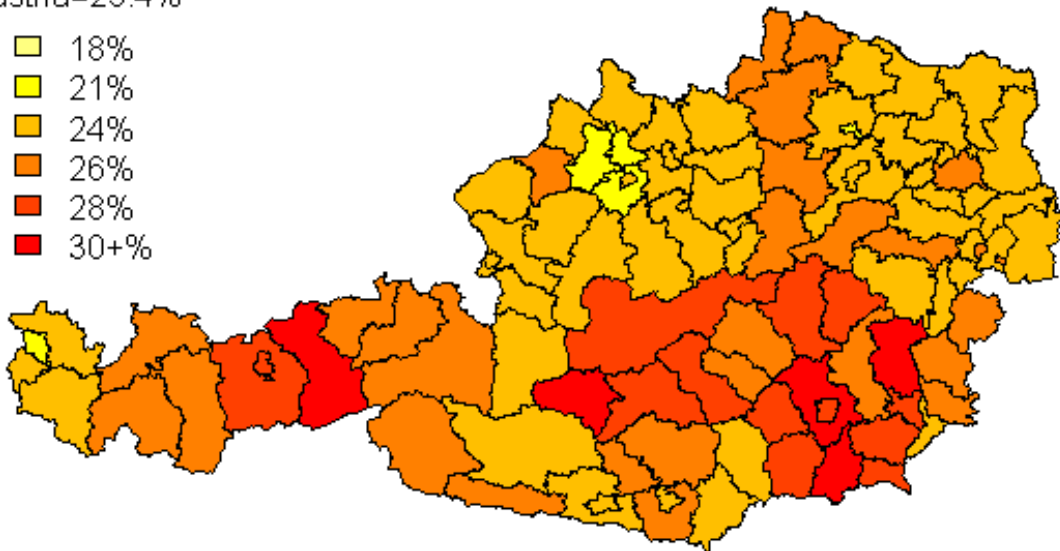
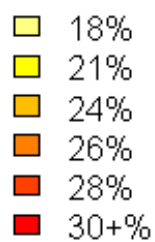
Very similar prevalence estimates were obtained for 2006. The overall and sex-specific CKD prevalences were 25.3% (SE 0.8%), males: 18.4% (1.0%), females: 32.0 (1.3%).

3.7 Regional differences in prevalence (Austria map)

The regional distribution of CKD prevalence in Austrian diabetic patients was age and sex standardized by the indirect method. As reference population, the full data set of diabetic patients in Austria was used.

Fig.7: Regional distribution of CKD prevalence in Austria.

Austria=25.4%



4 CONCLUDING REMARKS

By linking accurate laboratory measurements with drug prescriptions and hospital discharge records, we were able to estimate CKD rates even for persons for whom no laboratory measurements are available. The method does not even require that the subpopulation with laboratory measurements is unconditionally representative for the total population. We only assume that the high-dimensional prescription pattern is correlated with presence or stage of CKD. A similar assumption is also made in survey sampling methodology.

Having shown that prescription patterns highly correlate with the true CKD status of patient, the next logical step is to evaluate whether the high-dimensional drug prescription model may also be used as a diagnostic model for CKD, estimating the individual probability of presence of kidney damage. Such a model could guide general practitioners to refer or not refer their diabetic patients to laboratories and nephrologists. All necessary covariates are easy to collect, even more so with growing facilities to scan a patient's recently prescribed drugs from his/her health insurance card.

5 REFERENCES

Ambler G, Seaman S, Omar RZ. An evaluation of penalised survival methods for developing prognostic models with rare events. *Statistics in Medicine* 2012; 31 (11-12): 1150-1161.

Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. *International Journal of Epidemiology* 2002; 31(6): 1246-1252.

Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 2010; 33(1): 1-22.

Inskip H. Standardization methods. In: Gail MH, Benichou J, editors. *Encyclopedia of Epidemiologic Methods*. John Wiley & Sons 2000; 871–884.

Le Cessie S, Van Houwelingen JC. Ridge estimators in logistic regression. *Applied Statistics* 1992; 41(1): 191-201.

National Kidney Foundation. Kidney Disease Outcomes Quality Initiative. Clinical Practice Guidelines for Chronic Kidney Disease: evaluation, classification, and stratification. *American Journal of Kidney Diseases* 2002; 39(Suppl 1): s1-s266.

Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1996; 58(1): 267-288.