Beyond standardization: Reproducible approaches to deriving clinically meaningful variables for several measures of renal function

Amy Goodwin Davies,[1] Evanette Burrows,[1] Ashley Batugo,[1] Kimberley Dickinson,[1]
Sara Deakyne Davies,[2] Richard Hoyt,[3] Janet Zahner,[4]
Vikas R. Dharnidharka,[5] Bradley P. Dixon,[2] Joseph T. Flynn,[6]
Mark M. Mitsnefes,[4] William E. Smoyer,[3]
L. Charles Bailey,[1] Hanieh Razzaghi,[1] Michelle R. Denburg[1]

[1]Children's Hospital of Philadelphia, [2]Children's Hospital of Colorado, [3]Nationwide Children's Hospital, [4]Cincinnati Children's Hospital Medical, [5]St. Louis Children's Hospital, [6]Seattle Children's Hospital

Introduction

Standardization to a Common Data Model (CDM) unifies disparate data into a shared format[1]. For some measures, standardization is an initial step, followed by data cleaning and applying clinical understanding to derive additional variables, e.g., via thresholds for continuous data or calculations based on >1 existing variable. These steps should be incorporated into a reproducible workflow for systematic application in research or quality improvement analyses. This maximizes efficiency and avoids unwanted inconsistency across projects.

Our presentation will review our approach to several measures of renal function within PEDSnet[2], a network of children's hospital health systems which uses an expanded version of the OMOP CDM. Our team developed a series of reproducible approaches, implemented in R, which take CDM data as input and output clinically meaningful variables for levels of proteinuria, presence of hematuria, and estimated glomerular filtration rate (eGFR). This effort required clinical interpretation and thorough understanding of the data. Careful data quality investigation and remediation were also components; however, our focus is processing post-data-quality-remediation. In this abstract, we concentrate on proteinuria as a case study. Deriving proteinuria variables involved several components which were shared across our work on hematuria and eGFRs (table 1).

Table 1: Steps beyond standardization across three renal measures

	Proteinuria	Hematuria	eGFR
Classification of heterogenous result output	✓	✓	
Associating > 1 existing measure to derive additional measure	✓		✓
Data cleaning decisions	✓	✓	✓
Application of combined definition	✓	✓	

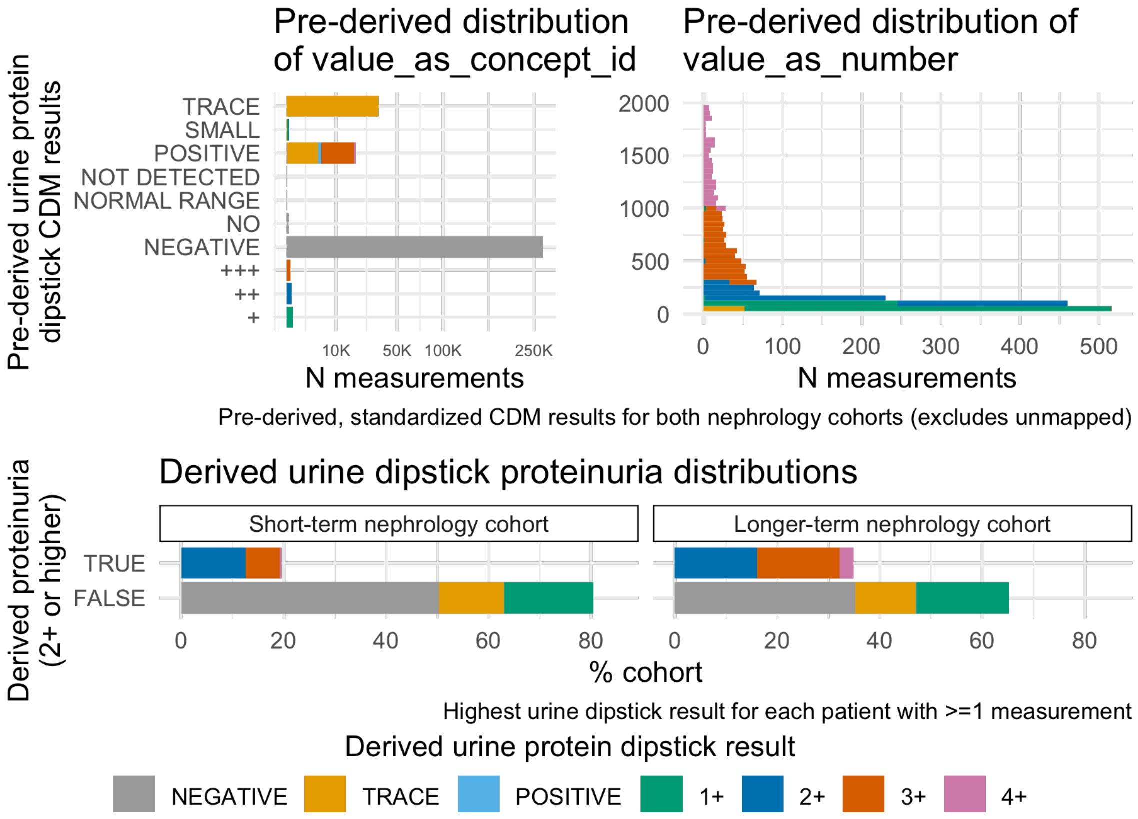
Methods & Results

To demonstrate the utility of the derived measures and provide some face-validity, we include distributions for 2 cohorts, "Short-term nephrology patients", patients with ≥1 nephrology encounter and "longer-term nephrology patients", patients with ≥2 nephrology encounters separated by ≥90 days (figure 1 and table 2). All patients were required to have ≥1 year of follow-up (2 face-to-face encounters of any specialty separated by ≥1 year). Data are from 6 institutions and PEDSnet database v4.1 (Jan. 2009 to Dec. 2020).

Proteinuria is measured in several distinct ways and representation in the PEDSnet CDM varies. Following clinician input, we focused on two measures: urine dipsticks (qualitative/semi-quantitative) and urine protein-to-creatinine ratios (UPCRs) (quantitative). UPCRs can be calculated from measurements of urine protein and urine creatinine concentrations when not reported directly.

For urine dipsticks, CDM results are heterogenous, reflecting variation in tests and source systems. For example, results can be represented on a “1+” to “4+” scale with value_as_concept_id or as numeric mg/dL estimates with value_as_number, with the following correspondence: Negative: <15, Trace: ≥15-<30, 1+: ≥30-<100, 2+: ≥100-<300, 3+: ≥300-<1000, 4+: ≥1000. It is necessary to derive additional variables which classify these results. Figure 1 illustrates how the heterogenous pre-derived CDM results were classified and the resulting distributions for derived proteinuria via urine dipstick for the 2 nephrology cohorts.

Figure 1: Distributions associated with deriving proteinuria via urine dipstick measurements



Pre-derived distribution of value_as_concept_id

Pre-derived distribution of value_as_number

Pre-derived, standardized CDM results for both nephrology cohorts (excludes unmapped)

Derived urine dipstick proteinuria distributions

Highest urine dipstick result for each patient with >=1 measurement

UPCRs are not always directly reported. To expand data capture for this variable we developed an approach which calculates UPCRs based on separate urine protein and urine creatinine measurements within a specified timeframe (i.e., 24 hours). Clinician engagement was required to define implausible values and determine sensible timeframes for associating measurements, based on lab ordering practices. This approach led to >2-fold increase in the number of patients with available data for both the short-term (5.17% → 12.5%) and longer-term (17.7% → 36.0%) nephrology cohorts (table 2).

UPCRs and urine dipstick measures can be incorporated into a combined measure to determine whether patients have any evidence for proteinuria. This can be defined as a urine dipstick measure of “2+” or above or a UPCR ≥ 0.2 mg:mg, as applied in table 2. 13.6% of the short-term cohort met this definition compared to 34.7% of the longer-term cohort.

Table 2: Demographic and clinical characteristics

Category	Characteristic	Short-term nephrology patients (N = 37,809)	Longer-term nephrology patients (N = 38,751)
Demographics and utilization	Follow-up (years, any specialty)	7.9 (4.3, 12.0)	7.6 (3.9, 12.1)
	Age at first visit (years)	2.6 (0.2, 8.3)	4.2 (0.3, 10.0)
	Female	17,820 (47.1%)	17,287 (44.6%)
	Encounters per person-year (any specialty)	3.3 (1.6, 6.5)	4.9 (2.6, 9.8)
	Nephrology encounters per person-year	0.2 (0.1, 0.3)	0.9 (0.4, 1.9)
Urine protein	Directly-reported UPCR available	1,954 (5.2%)	6,850 (17.7%)
	Directly-reported UPCR	0.14 (0.07, 0.40)	0.22 (0.09, 0.80)
	Proteinuria via directly-reported UPCR (≥0.2 mg:mg)		
	within entire cohort	861 (2.3%)	4,230 (10.9%)
	within cohort with available measurement(s)	861 (44.1%)	4,230 (61.8%)
	Derived UPCR available	4,733 (12.5%)	13,933 (36.0%)
	Derived UPCR	0.15 (0.07, 0.42)	0.21 (0.09, 0.76)
	Proteinuria via derived UPCR (≥0.2 mg:mg)		
	within entire cohort	2,277 (6.0%)	9,094 (23.5%)
	within cohort with available measurement(s)	2,277 (48.1%)	9,094 (65.3%)
	Urine dipstick protein measurement available	20,044 (53.0%)	28,823 (74.4%)
	Proteinuria via urine dipstick (≥2+)		
	within entire cohort	3,924 (10.4%)	10,034 (25.9%)
	within cohort with available measurement(s)	3,924 (19.6%)	10,034 (34.8%)
	Proteinuria via combined definition (≥2+ or ≥0.2 mg:mg)		
	within entire cohort	5,153 (13.6%)	13,429 (34.7%)
	within cohort with available measurement(s)	5,153 (24.3%)	13,429 (42.9%)
Other measures of renal function	Urine blood measurement available	21,215 (56.1%)	29,607 (76.4%)
	Hematuria		
	within entire cohort	7,480 (19.8%)	14,675 (37.9%)
	within cohort with available measurement(s)	7,480 (35.3%)	14,675 (49.6%)
	eGFR available	21,710 (57.4%)	32,453 (83.8%)
	eGFR	106.2 (88.4, 125.9)	100.1 (79.8, 119.8)

Categorical reported as N patients (% cohort) and continuous reported as median (IQR)
Results calculated across all available data for patients

Discussion

As the case study of proteinuria illustrates, without taking additional processing steps beyond the standardized CDM data, results can be heterogenous and challenging to interpret or analyze (e.g., urine dipsticks) or unavailable for many patients (e.g., UPCRs). A reproducible approach to these additional steps maximizes efficiency and avoids unwanted inconsistency across projects. However, some decisions will be made at a project-specific level. Ideally, we want to allow flexibility, whilst clearly documenting decisions and avoiding unnecessary variation. Ongoing development aims to constrain variation in down-stream processing by parameterizing the functions for project-specific decisions.

The results presented for the 2 nephrology cohorts (table 2) provide some face-validity for our measures of proteinuria, hematuria, and eGFR. As expected, higher proportions of patients in the longer-term nephrology cohort have urine protein, urine blood, and eGFR measurements available, compared to the short-term nephrology cohort. Furthermore, higher proportions of these patients meet definitions for proteinuria and hematuria.

A limitation of our general approach is it requires manual review of result distributions, so periodic re-review will be required as data is updated. Manual review has implications for work conducted in a distributed network, as we would want to examine distributions before executing the final query (this could be achieved through a preliminary query which returns result distributions).

Conclusion

For some measures, standardization is an initial step, which must be followed by data cleaning and applying clinical understanding to derive additional variables. Our team developed a series of reproducible approaches, which take CDM data as input and output analysis-ready, clinically meaningful variables for levels of proteinuria, presence of hematuria, and eGFR. These approaches, implemented in R, will be made publicly available to accompany our presentation.

References

1. Observational Health Data Sciences and Informatics. The Book of OHDSI. 1st ed. 2021 Available from: https://ohdsi.github.io/TheBookOfOhdsi/ [Accessed 17th June 2021].
2. Forrest C, Margolis P, Bailey LC, Marsolo K, Del Beccaro M, …, Kahn M. PEDSnet: a national pediatric learning health system. J Am Med Inform Assoc. 2014; 21(4):602-606.