Supplementary Data for

Laajala TD, Murtojärvi M, Virkki A, Aittokallio T. *ePCR*: an R-package for survival and time-to-event prediction in advanced prostate cancer, applied to real-world patient cohorts.

Real-world registry data processing

A notification of the registry-based study design was made to the Office of the Data Protection Ombudsman according to the appropriate legislation, and the data gathering and analysis was performed with the study permission of Varsinais-Suomen sairaanhoitopiirin kuntayhtymä (approval T287/2016). The patient registry data were provided by the Turku University Hospital Centre for Clinical Informatics in an SQL database. The database contains information of the prostate cancer patients treated in the hospital, mainly during the years 2004-2016, although some diagnoses were also present from the years 2002-2003. There were altogether 8493 patients, of which 7997 had a diagnosis of prostate cancer, identified with ICD10 code C61 or C61&. Information on prescription medications was mainly available since the mid-2010. Regular expressions were found to be sufficient for extracting the model variables, originally derived from the US-based clinical trials. Prescription medications were matched based on their ATC codes, and clinical diagnoses based on ICD 10 codes. Medical operations, such as surgeries, and the values for the clinical measurements were identified by national and hospital-specific codes[1]. For routine clinical measurements, such as BMI, height, weight and blood pressure, variable matching was based on their Finnish labels. Except for patient cohort selection (see below), only the structured data was used in these analyses, that is, we did not extract information from the patient medical records that were available as free-form text only.

Patient cohort selection criteria

Two sub-cohorts of advanced prostate cancer patients were extracted from the pool of all the available prostate cancer patients in the real-world registry data (n=7997).

The first cohort was identified using selection criteria similar to Seyednasrollah et al (2017). We tried to replicate as closely as possible their analyses, in terms of the clinical variables and modelling choices, as well as the reported patient selection criteria using medication information only (personal communication and the supplement of Seyednasrollah et al., 2017). More precisely, all the diagnosed prostate cancer patients were included that had been treated with both antiandrogens (ATC codes L02BB*, G03H*, where * means that anything is allowed) and docetaxel. Docetaxel treatments were identified using all common Finnish spellings of the treatment name, as ATC codes were not available for the treatments given at the hospital. Different from Seyednasrollah et al. (2017), we did not implement further exclusion criteria using clinical trial eligibility status or other malignancies, to extract as large cohort as possible. 180 patients matching these selection criteria were identified, referred to as "***patient selection by medication only***".

The second cohort was found by text mining of the patient medical records, referred to as "***patient selection by medical records***". The implemented algorithm searched for sentences containing phrases, such as "castration resistant" or "hormone refractory", which did not include words indicating uncertainty, such as "suspicion" or "is developing"[2]. Common spelling errors were also accepted in the search process. The

[1] These codes are available at http://www.terveysportti.fi/terveysportti/toimenpideluokitus.koti and https://webohjekirja.mylabservices.fi/TYKS/ (in Finnish)

[2] The medical records were in Finnish and the exact regular expressions were: one of '.*kastraatioresist', '.*crpc', '.*hormoniresist', '.*hormooniresist', '.*gastraatioresist', '.*hormonirefr', '.*hormoonirefr', '.*kastraatiorefr',

vocabulary for the text search was iteratively improved manually by trial-and-error. An evaluation by a clinical expert found that less than 2% of the CRPC classifications were erroneous, while ca. 15% were uncertain. The text search resulted in a larger cohort of 587 patients. Since we only looked for castration resistance, and did not explicitly include criteria for metastases, this cohort was larger but potentially more heterogeneous in terms of the advanced stage of the disease, in comparison to the first cohort.

Survival analyses

Survival and time-to-event prediction was based on ensemble of penalized Cox regression (ePCR) model (Guinney et al., 2017). The start time was selected based on the first time of the docetaxel treatment in the medication-selected cohort, and on the first recorded identification of castration resistance in the second cohort. The time-to-event was computed as the difference (in days) between the patient-specific start time and death (observed end-point event), or the last observation of the patient (censoring). This resulted in a right-censored observation vector, which is typical for survival analysis, consisting of two components: the day to event or censoring and an indicator for event or censoring. Clinical features for the prognostic model included the latest eligible observations dated before or on the start day of the given patient.

Prognostic modelling

Four prognostic models were included in the analyses: the Halabi model (Halabi et al., 2014), the winning ePCR model of the Prostate Cancer DREAM Challenge (Guinney et al., 2017), a reduced version of the full ePCR model (see below), and the version of the original ePCR model used by Seyednasrollah et al. (2017). A clinical variable of the full ePCR model was included in the reduced model only if there were less than 40% of missing values of the particular variable in the patient cohort identified by text searches of the medical records. Here, a value was considered missing for a patient if it had never been measured for the patient before the time of the diagnosis (start time). Since the clinical measurements were not always provided using the same unit, the pint library[3] was used for automatically converting the clinical features to the same units as in the DREAM clinical trial data. The cases where the conversion failed were corrected manually. Finally, the values of certain features were log-transformed. The exact variables used by Seyednasrollah et al. (2017) remained somewhat uncertain since some variables were not specifically reported in their original work. The feature abbreviations in Suppl. Table 1 are as presented in the DREAM Challenge data and their full descriptions can be found from the Data Dictionary accompanying Guinney et al. (2017).

For comparison of our results to those of Seyednasrollah et al. (2017), an additional four-week limit on the time period of observations (based on the start time of each patient) was imposed, similarly as in Seyednasrollah et al. (2017). The ePCR models were also tested without the time limit, which generally improved the prediction performances (see Figure 1 in the manuscript). When no measurement for a clinical variable was available, it was replaced by an imputed value. Two imputation methods were tested for the real-world hospital patient cohort: median imputation and *k*-NN imputation, with default $k = 10$. In both cases, only the information present in the variable table for the patient cohort was taken into account. For two of the variables of the full ePCR model (urine specific gravity SPEGRA and patient performance status ECOG_C), there were no observations in the raw data tables. These missing values were replaced by a constant value (the median of the same variable in the DREAM Challenge clinical trial data). We note that these variables were not included in our reduced ePCR model (see Suppl. Table 1).

'.*gastraatiorefr' and none of '.*kehittymässä', '.*muuttumassa', '.*epäily', '.*mikäli', '.*kehittyy', '.*jos[\s,\.]', '.*vaikuttaa'.

[3] https://pint.readthedocs.io/en/latest/

The patient registry data was used for selecting the subset of clinical variables available in the real-world cohorts and included in the reduced model (60 variables that were available for at least 60% of patients in the patient cohort selected by medical records), but this selection was done based on the missing value rate only, without using the prediction accuracy as a selection criterion (i.e., not leading to over-fitting). This is because in a real predictive setting with newly-hospitalized patients, the survival of those patients is obviously unknown and cannot therefore be used as a criterion for selecting the model variables. In Fig. 1A, however, the rightmost bars were obtained by testing two models and selected best ways to pre-process the real-word data in the same cohort used for evaluating the prediction accuracy. In particular, we tested the effect of imposing a time limit on the observations (similar to Seyednasrollah et al, 2017), as well as different imputation methods (K-NN and median imputation). In real applications, one could not expect to obtain as good results as in the best cases of Fig. 1A, where survival information was used for selecting the best model, and therefore these particular results are expected to be optimistically biased to some degree.

Ensemble model objects available

The readily-fitted ensemble components for the DREAM clinical trial cohort from Guinney et al. (2017) and the ensemble combining the two real-world patient cohorts from Turku University Hospital (e.g. regarding the selected variables and their estimated penalized coefficients) are provided as-is in the ePCR R-package. As such, the provided ensembles can be used for model inference and future predictions, but the S4-class R-object of the original data matrix has been omitted, due to personal data and security reasons. The original clinical data matrices can be obtainable from their respective owners (Project Data Sphere, https://www.projectdatasphere.org; and the Turku University Hospital, Centre for Clinical Informatics http://cci.vsshp.fi with an appropriate study permit). The ensemble model for the synthetic patient data simulated based on the Turku University Hospital cohort (see below) is available as an ePCR-package S4-object. All the R-functions for the clinically relevant prediction tasks, including overall survival and time-to-event prediction for future data, as well as incorporated analytics and diagnostic plots for the penalized Cox regression models, can be run on model objects without including the original data matrices. The prediction accuracy was evaluated using the c-index and iAUC functions in the ePCR-package; for more detailed instructions, please see the reference manual (https://cran.r-project.org/web/packages/ePCR/ePCR.pdf).

Generating simulated patient data

Due to the patient data security and confidentiality issues, the exact patient cohort measurements cannot be embedded to the open-source, freely available R-package itself. For simulating realistic, yet synthetic data, a nonparametric mixture kernel density function was estimated based on the real-world patient data using the *np*-package in R (Hayfield & Racine, 2008), after which simulated observations were drawn from the estimated mixture probability density functions. As bandwidth length is a key aspect in nonparametric kernel density estimation in the *np*-package, our simulation approach made use of computationally intensive cross-validation for selecting optimal bandwidths for each dimension. In order to avoid getting stuck at local minima, the adaptive bandwidth selection procedure was run with multi-start mode, i.e., varying multiple starting parameters were used for the optimization procedure. For continuous numeric variables, conventional second-order Gaussian kernels were used. For non-continuous clinical variables, the *np*-package offers the possibility to take into account the mixture nature of clinical data regarding unordered (binary or nominal) and ordered (ordinal) categorical features, for which the default kernels (Aitchison-Aitken and Wang-van Ryzin, respectively) were used. These latter, non-conventional kernels for categorical features are documented in the corresponding R-package publication (Hayfield & Racine, 2008).

After estimating the kernel density probability density function (PDF), an importance sampling procedure was adopted to generating synthetic patients by sampling from the true observed values of each feature independently. A large number of such synthetic patients were generated (>100k), thus approximating the full spectrum of possible patient characteristic combinations in the high-dimensional space (>100 features). A smaller sample of patients from this pool of candidates was randomly chosen as the final synthetic dataset by sampling with weighting by the corresponding kernel-based PDF. Furthermore, a representative survival response vector was generated from the Weibull distributions with censoring characteristics similar to the true observed event and censoring time quantiles. More specifically, 150 and 500 synthetic patients were sampled from the kernel density estimated PDFs for the two sub-cohorts of the Turku University Hospital registry data, with the original patient sample sizes of 180 and 587, respectively. Based on the exploratory diagnostics for PCA and boxplots (PCA shown for various subsets and in relation to the DREAM cohort in Supplementary Figures 1A-D), the key distributional characteristics were well-captured for all the mixture data types: continuous, ordered variables and unordered variables. The representativeness of the simulated response vectors were verified using the Kaplan-Meier plots (Supplementary Figures 1E-F). These simulated data matrices and the corresponding response vectors are provided in the *ePCR*-package for illustrating the use of the package, as well as to provide representative data emerging from the real-world hospital patient cohorts. The kernel density-simulated data captured the relevant data characteristics also based on visual diagnostics for the univariate continuous or binary/ordinal features in Supplementary Figure 2 and Supplementary Figure 3, respectively.

Supplementary references

Justin Guinney, Tao Wang, Teemu D Laajala, et al. Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: development of a prognostic model through a crowdsourced challenge with open clinical trial data. The Lancet Oncology 2017; 18 (1): 132- 142.

Susan Halabi, Chen-Yen Lin, W. Kevin Kelly, Karim S. Fizazi, Judd W. Moul, Ellen B. Kaplan, Michael J. Morris, and Eric J. Small. Updated Prognostic Model for Predicting Overall Survival in First-Line Chemotherapy for Patients With Metastatic Castration-Resistant Prostate Cancer. Journal of Clinical Oncology 2014; 32(7): 671-677.

Fatemeh Seyednasrollah, Mehrad Mahmoudian, Liisa Rautakorpi, Outi Hirvonen, Tarja Laitinen, Sirkku Jyrkkiö, Laura L. Elo. How Reliable are Trial-based Prognostic Models in Real-world Patients with Metastatic Castration-resistant Prostate Cancer? European Urology 2017; 71(5): 838-840.

Jerome Friedman, Trevor Hastie, Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 2010; 33(1): 1–22.

Tristen Hayfield, Jeffrey S Racine. Nonparametric Econometrics: The np package. Journal of Statistical Software 2008; 27(5): 1-32.

Supplementary Tables

Supplementary Table 1. Model variables in the four prognostic models used in the analyses.

Variable	Unit	Full ePCR	Seyednasrollah et al.	Reduced ePCR	Halabi et al.
BMI	kg/m^2	X	X		
HEIGHTBL	cm	X	X		
WEIGHTBL	kg	X	X	X	
ALP*	U/l	X	X	X	X
ALT*	U/l	X	X	X	
AST*	U/l	X	X		
CA	mmol/l	X	X		
CREAT*	µmol/l	X	X	X	
HB	g/dl	X	X	X	X
LDH*	U/l	X			X
NEU*	10^9/l	X	X		
PLT	10^9/l	X	X	X	
PSA*	ng/ml	X	X	X	X
TBILI*	µmol/l	X	X		
TESTO*	nmol/l	X	X		
WBC*	10^9/l	X	X	X	
CREACL*	ml/min	X			
NA.	mmol/l	X	X	X	
MG*	mmol/l	X			
PHOS*	mmol/l	X			
ALB	g/l	X	X		X
TPRO	g/l	X	X		
RBC	10^{12}/l	X	X	X	
LYM*	10^9/l	X			
BUN*	mmol/l	X			
CCRC*	ml/min	X			
GLU*	mmol/l	X			
SYSTOLICBP	mmHg	X	X		
DIASTOLICBP	mmHg	X	X		
PULSE	bpm	X	X		
HEMAT	%	X	X	X	
SPEGRA	kg/l	X	X		
LYMperLEU	%	X			
MONO	10^9/l	X	X		
MONOperLEU	%	X	X		
NEUperLEU	%	X	X		
POT	mmol/l	X	X	X	
BASOperLEU	%	X	X		

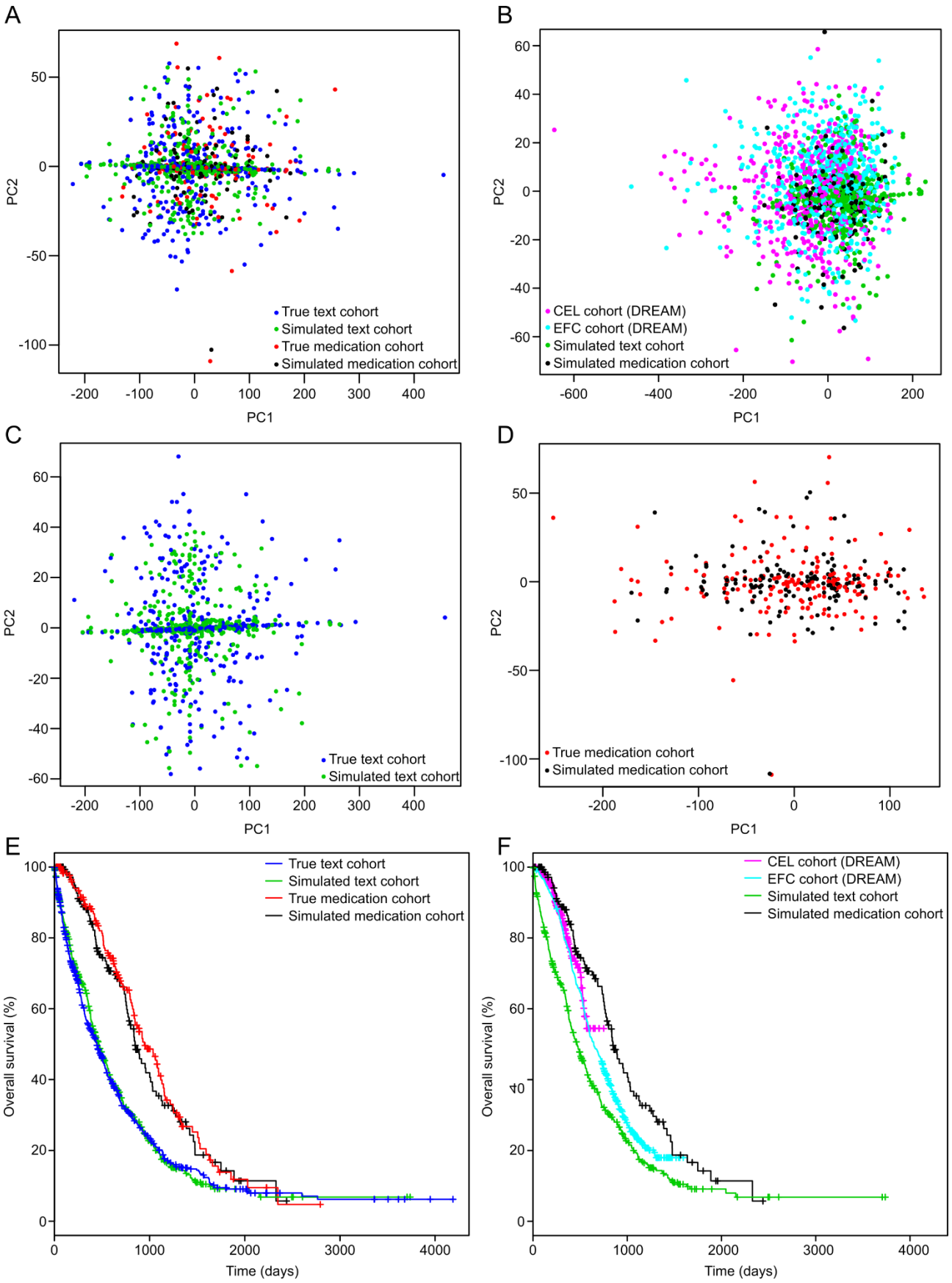
EOS	10^9/l	X	X		
EOSperLEU	%	X	X		
TARGET		X	X		
LYMPH_NODES		X	X	X	
KIDNEYS		X	X	X	
LUNGS		X	X	X	
LIVER		X	X	X	
PLEURA		X	X	X	
OTHER		X	X	X	
PROSTATE		X	X		
ORCHIDECTOMY		X	X	X	
PROSTATECTOMY		X	X	X	
LYMPHADENECTOMY		X	X	X	
BILATERAL_ORCHIDECTOMY		X	X	X	
PRIOR_RADIOTHERAPY		X	X	X	
ANALGESICS		X	X	X	X
ANTI_ANDROGENS		X	X	X	
GLUCOCORTICOID		X	X	X	
GONADOTROPIN		X	X	X	
BISPHOSPHONATE		X	X	X	
CORTICOSTEROID		X	X	X	
IMIDAZOLE		X	X	X	
ACE_INHIBITORS		X	X	X	
BETA_BLOCKING		X	X	X	
HMG_COA_REDUCT		X	X	X	
ESTROGENS		X	X	X	
ANTI_ESTROGENS		X	X	X	
CEREBACC		X		X	
CHF		X		X	
DVT		X		X	
DIAB		X		X	
MI		X		X	
PULMEMB		X		X	
SPINCOMP		X		X	
COPD		X		X	
MHBLOOD		X	X	X	
MHCARD		X	X	X	
MHCONGEN		X	X	X	
MHEAR		X	X	X	
MHENDO		X	X	X	
MHGASTRO		X	X	X	
MHHEPATO		X	X	X	
MHIMMUNE		X	X	X	
MHINFECT		X	X	X	

MHINJURY		X		X	
MHINVEST		X			
MHMETAB		X	X	X	
MHPSYCH		X	X	X	
MHRENAL		X	X	X	
MHRESP		X	X	X	
MHSKIN		X	X	X	
MHVASC		X		X	
ECOG_C		X	X		X
AGEGRP2		X	X	X	
RaceAsian		X	X		
RaceBlack		X	X		
RaceOther		X	X		
RaceWhite		X	X		
RegionAsia		X	X		
RegionEastEuro		X	X		
RegionNorthAmer		X	X		
RegionSouthAmer		X	X		
RegionWestEuro		X	X		
Disease site[4]					X

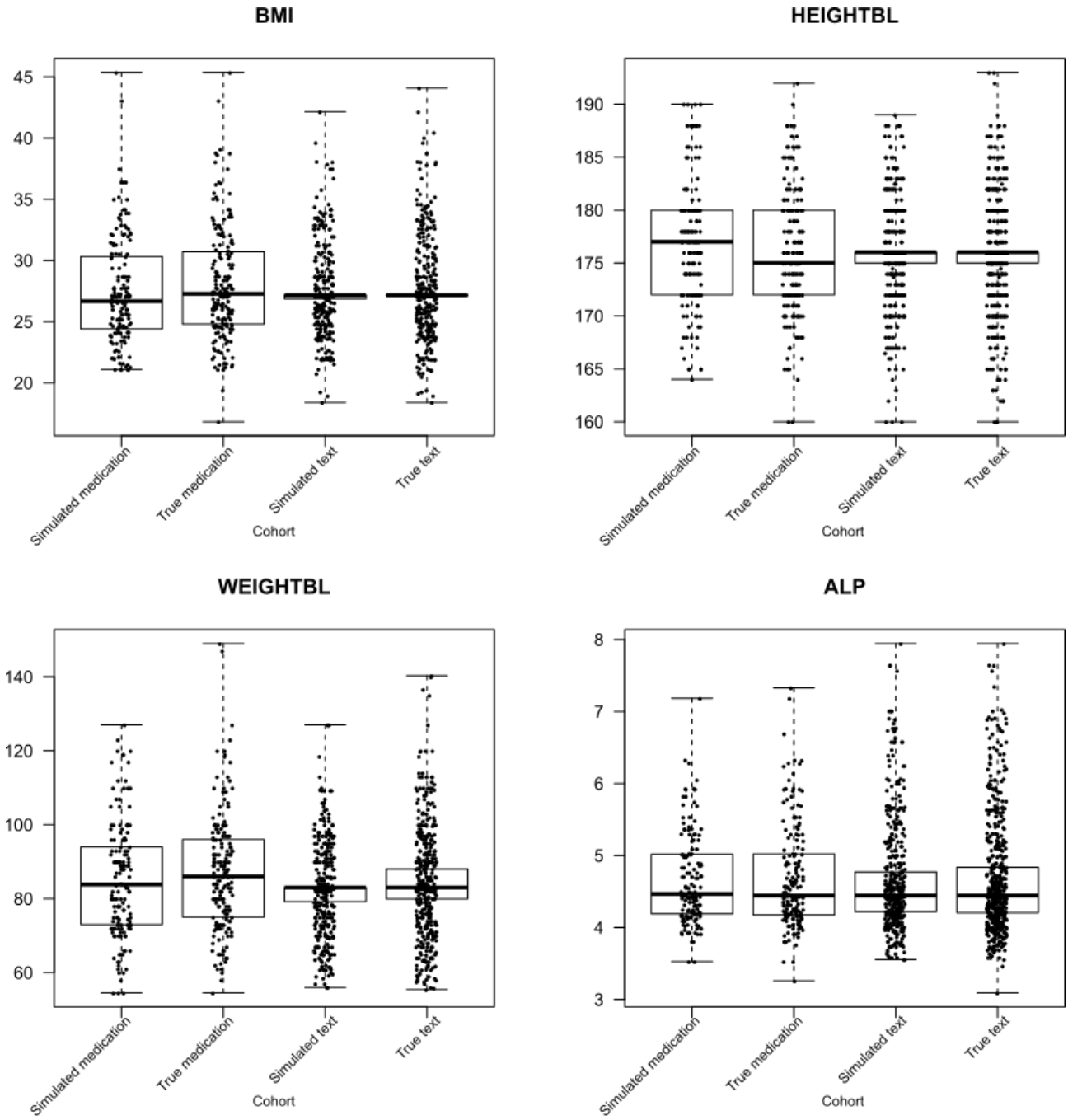
* Log-transformed variables.

[4] The variable ”Disease site” in the Halabi model indicates the site of metastases (lymph nodes, bone or visceral). The same information is used also in the ePCR models but represented as separate binary variables.

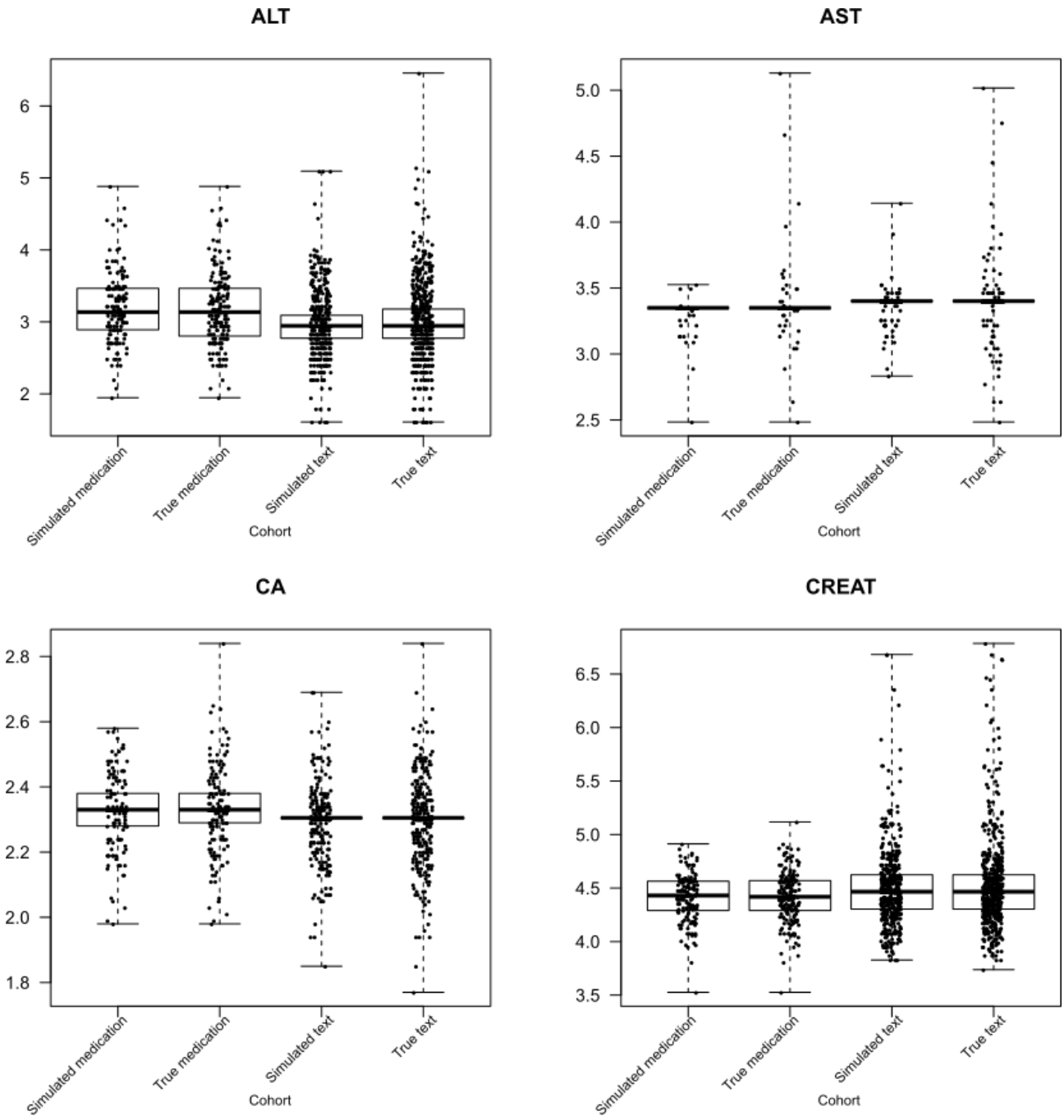
Supplementary Figures



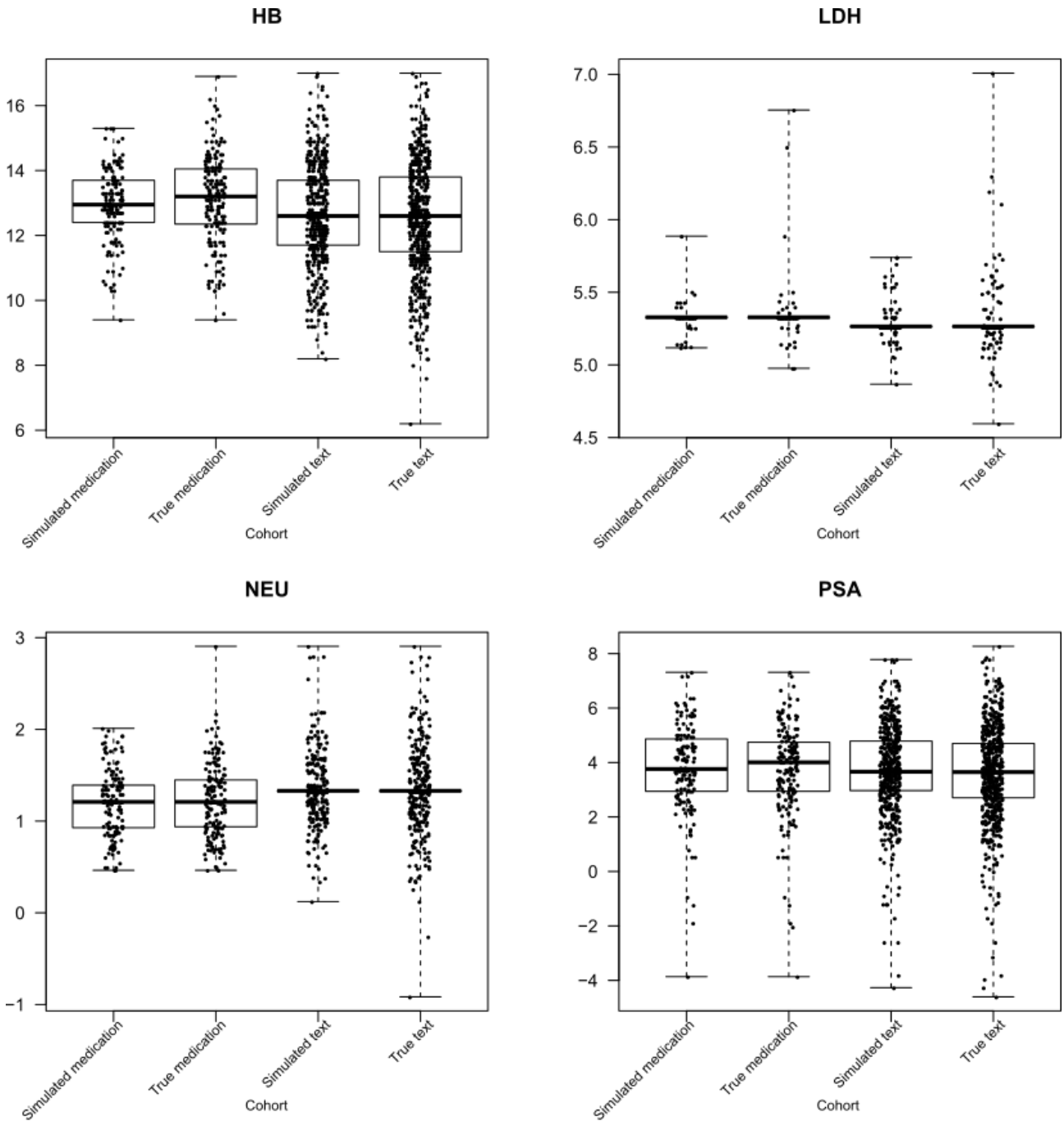
Supplementary Figure 1. Comparison of the simulated data estimated based on Turku University Hospital cohort with the true patient registry data, as well as with the DREAM Challenge 9.5 mCRPC cohorts utilized in the original top-performing ePCR model. All the 101 variables of the full ePCR model were used for generating the figures (see Supplementary Table 1). (**A**) PCA plot of simulated real-world data versus true data; (**B**) PCA plot of simulated text search-based patient cohort versus the true text search-based cohort; (**C**) PCA plot of simulated medication-based cohort versus the true medication-based cohort; (**D**) PCA plot of simulated real-world data versus two DREAM clinical trial cohorts; (**E**) Kaplan-Meier survival plots for the simulated and true real-world cohorts based on the simulated response vector; (**F**) Kaplan-Meier survival plots for the simulated real-world cohorts as well as the DREAM clinical trial training cohorts.



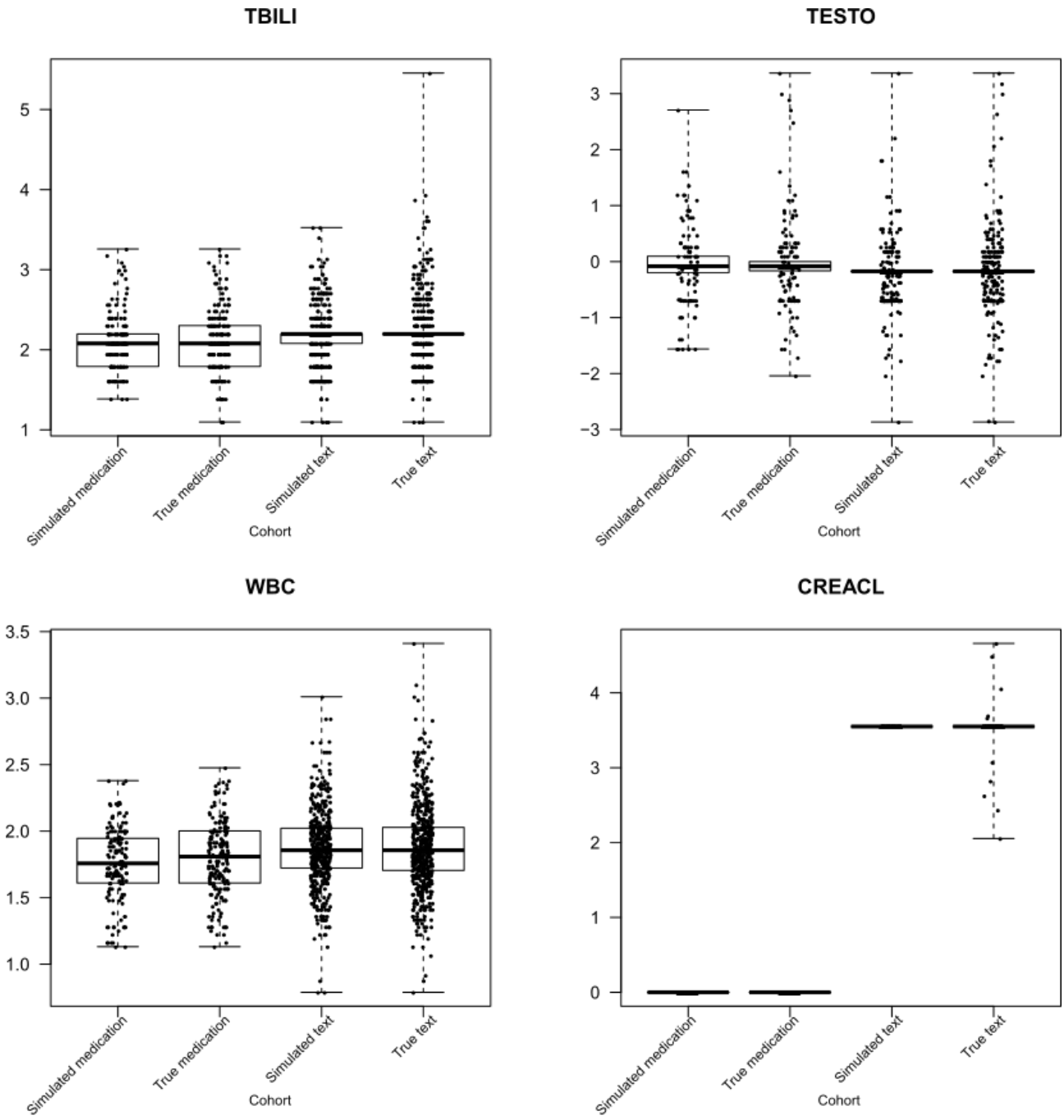
Supplementary Figure 2. Diagnostic boxplots for numeric variables in the simulated and true medication and text search-based cohorts.



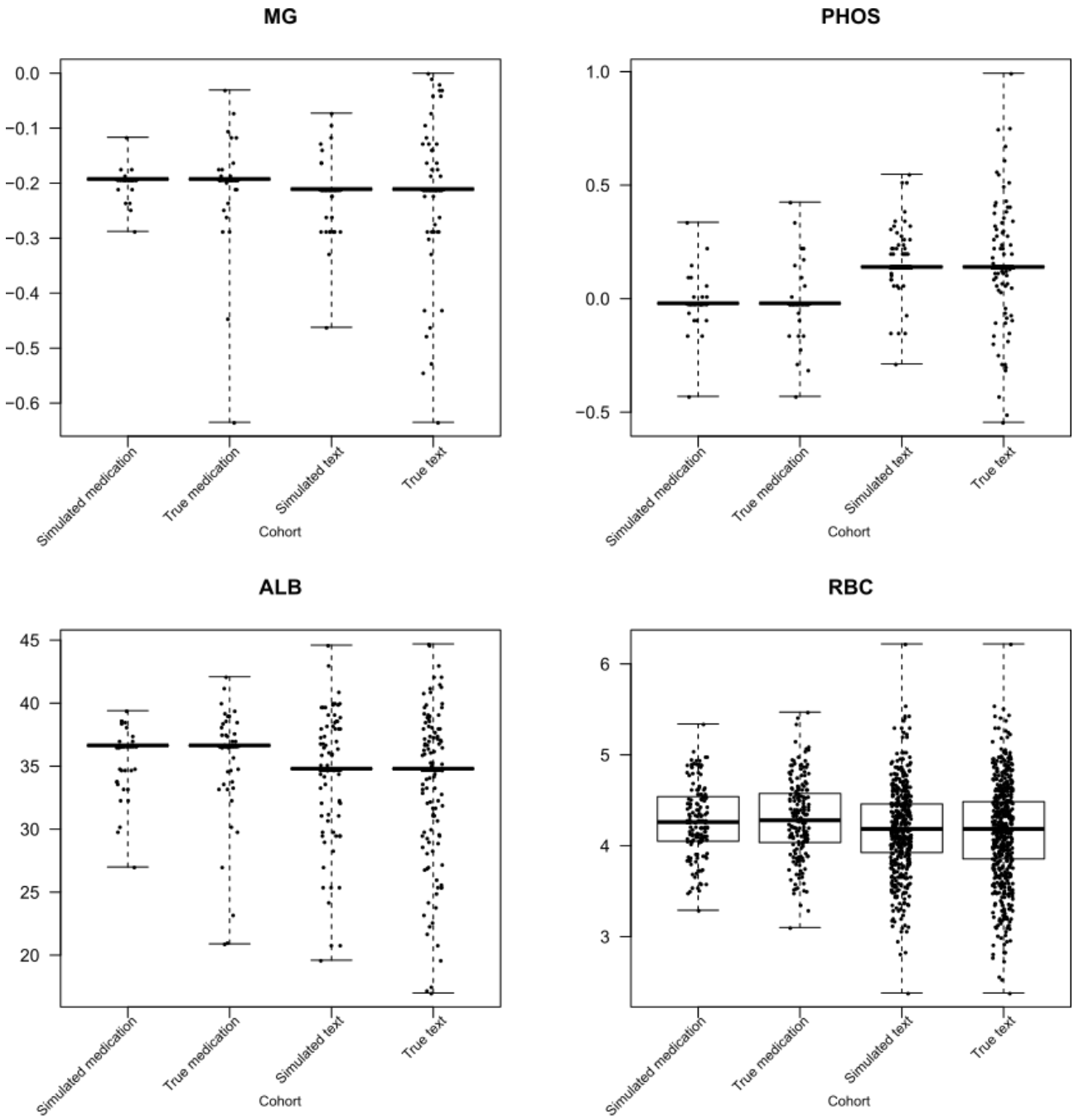
Supplementary Figure 2 (continued). Diagnostic boxplots for numeric variables in the simulated and true medication and text search-based cohorts.



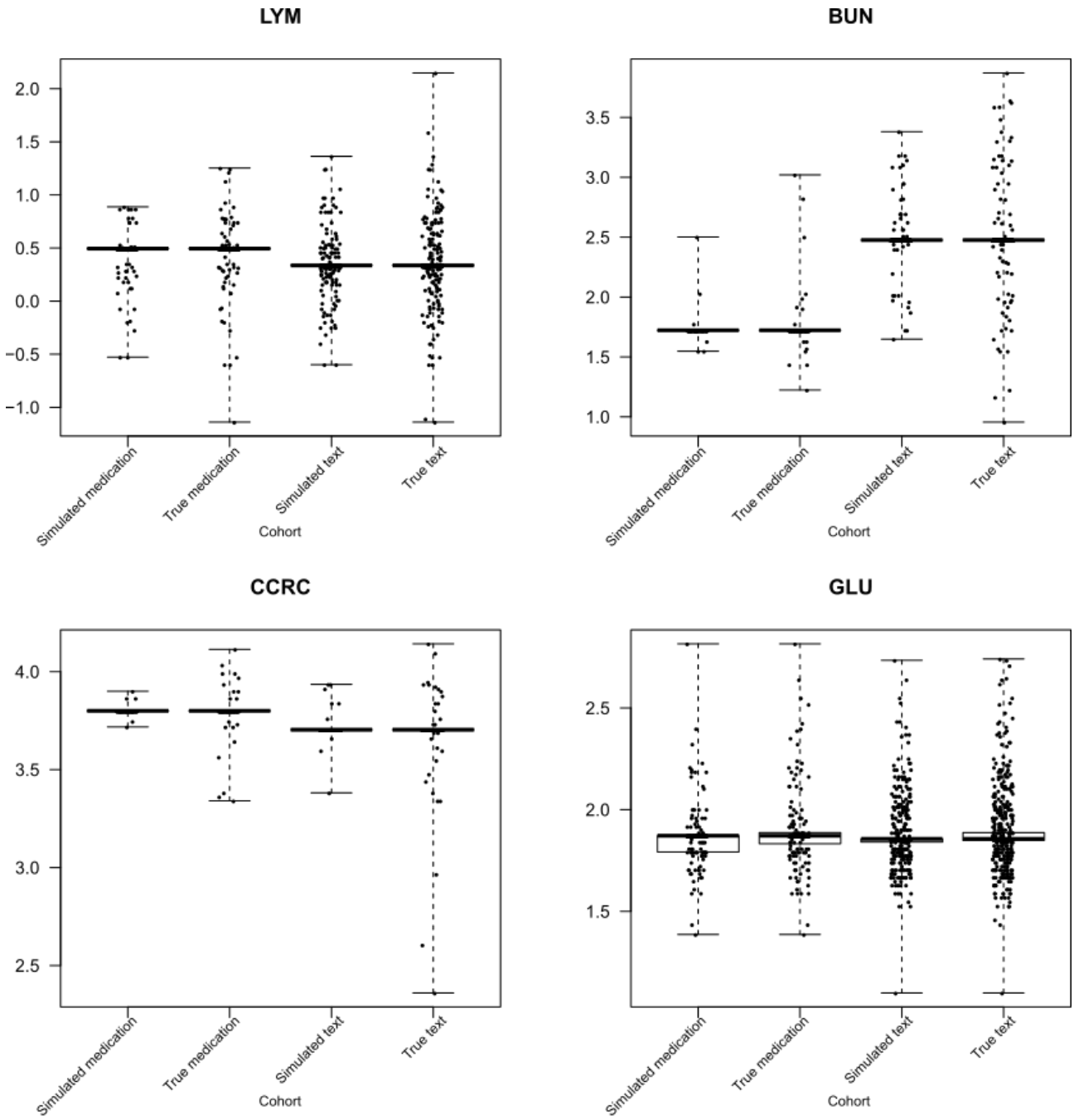
Supplementary Figure 2 (continued). Diagnostic boxplots for numeric variables in the simulated and true medication and text search-based cohorts.



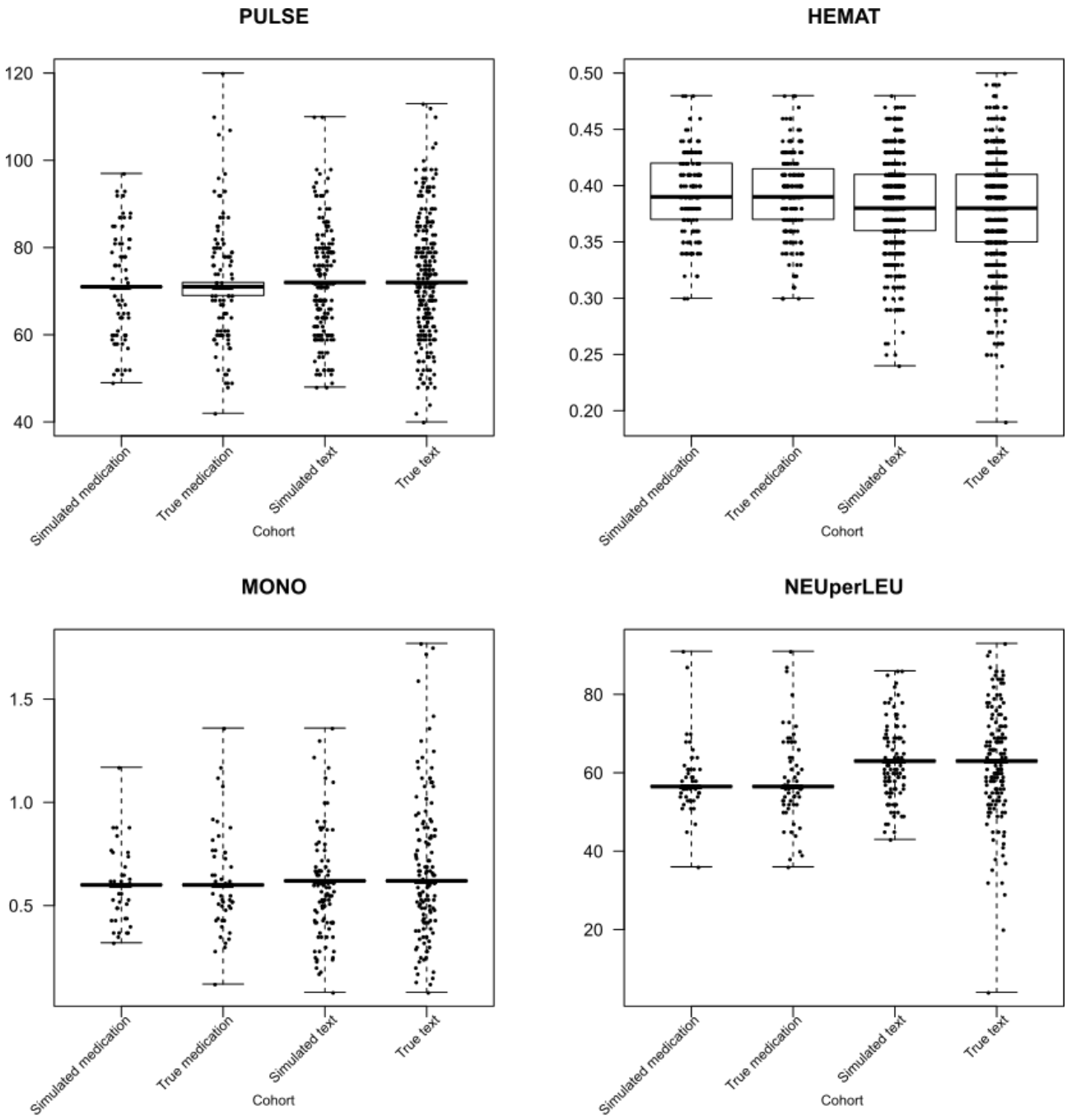
Supplementary Figure 2 (continued). Diagnostic boxplots for numeric variables in the simulated and true medication and text search-based cohorts.



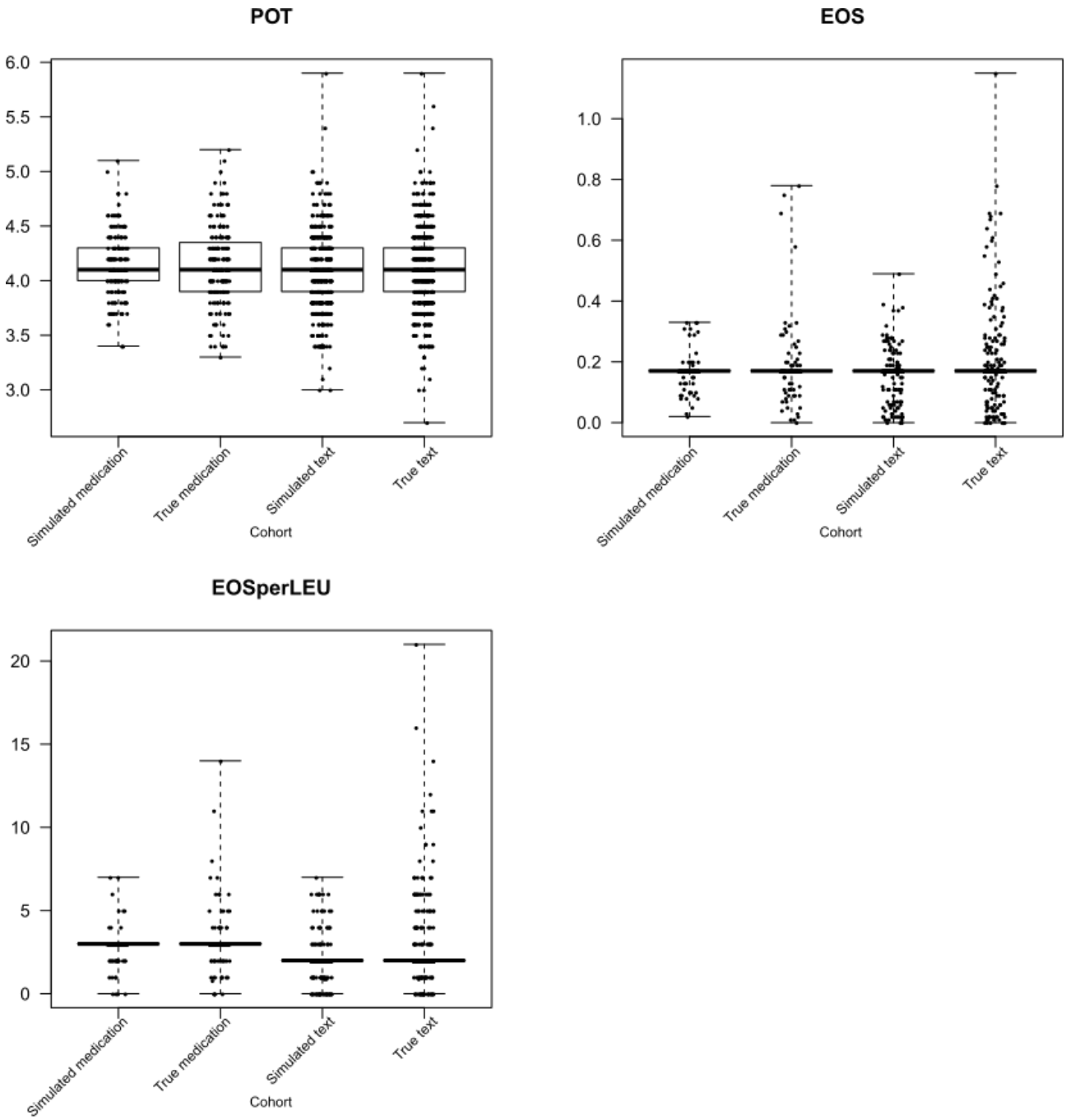
Supplementary Figure 2 (continued). Diagnostic boxplots for numeric variables in the simulated and true medication and text search-based cohorts.



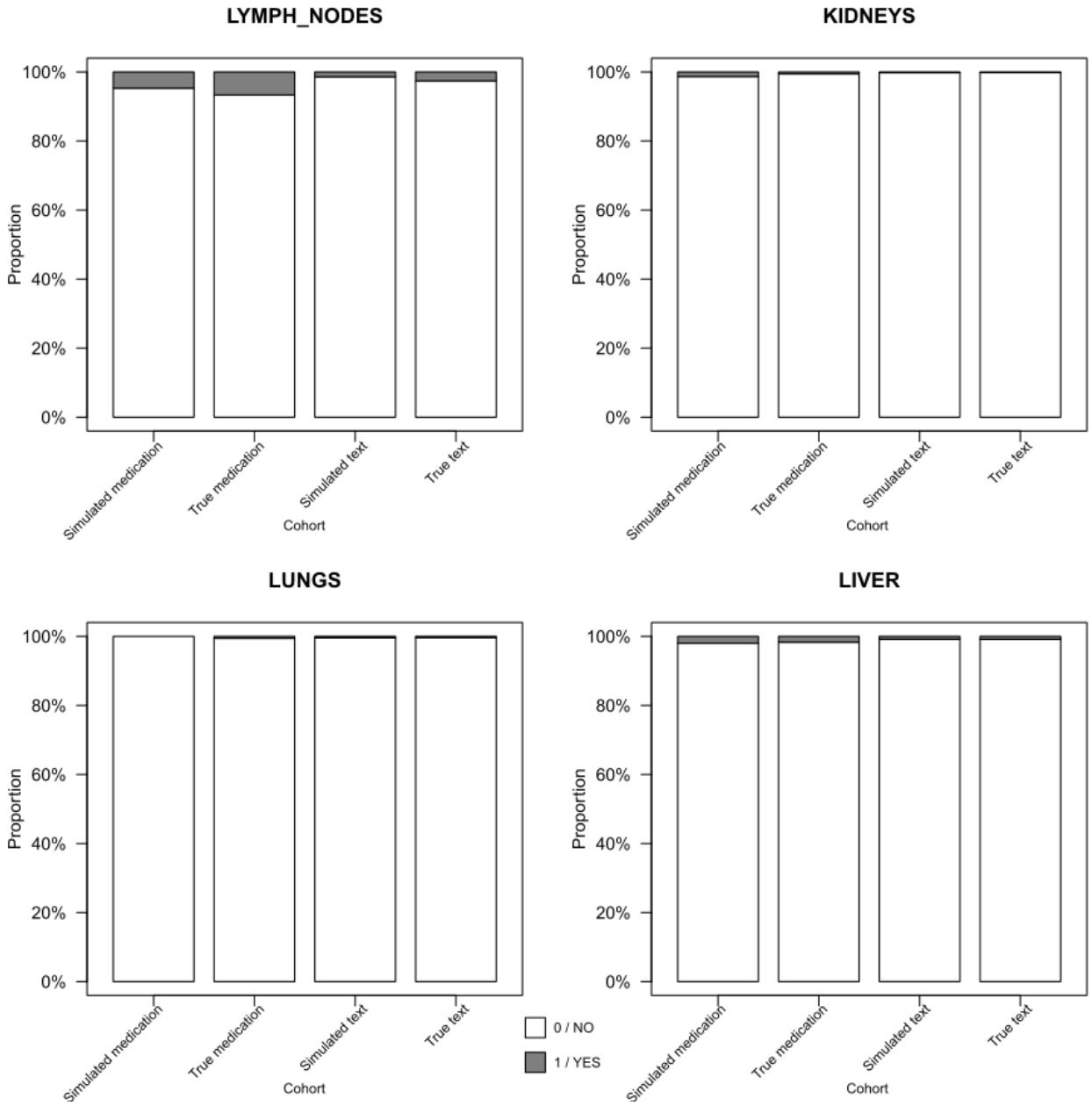
Supplementary Figure 2 (continued). Diagnostic boxplots for numeric variables in the simulated and true medication and text search-based cohorts.



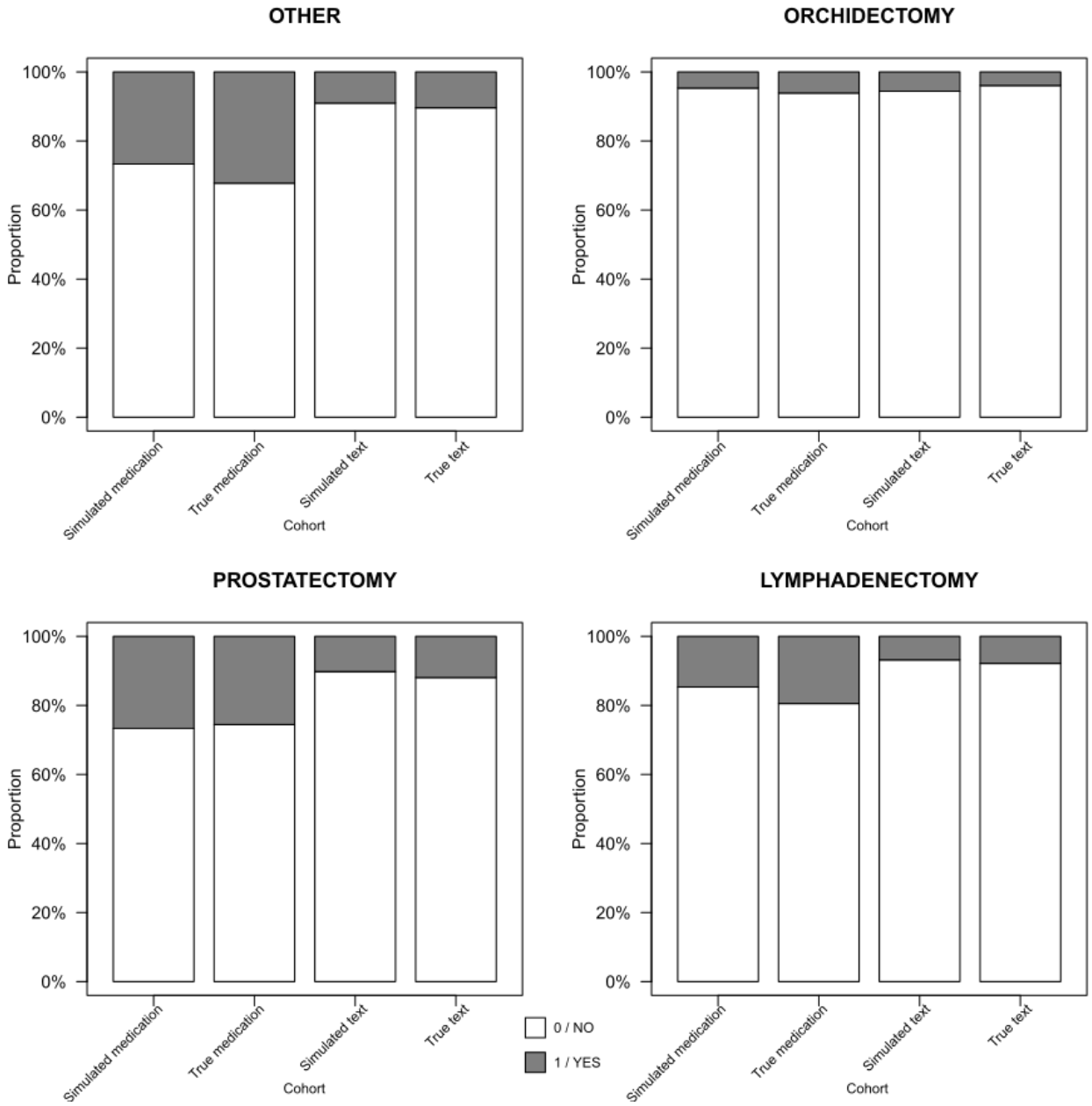
Supplementary Figure 2 (continued). Diagnostic boxplots for numeric variables in the simulated and true medication and text search-based cohorts.



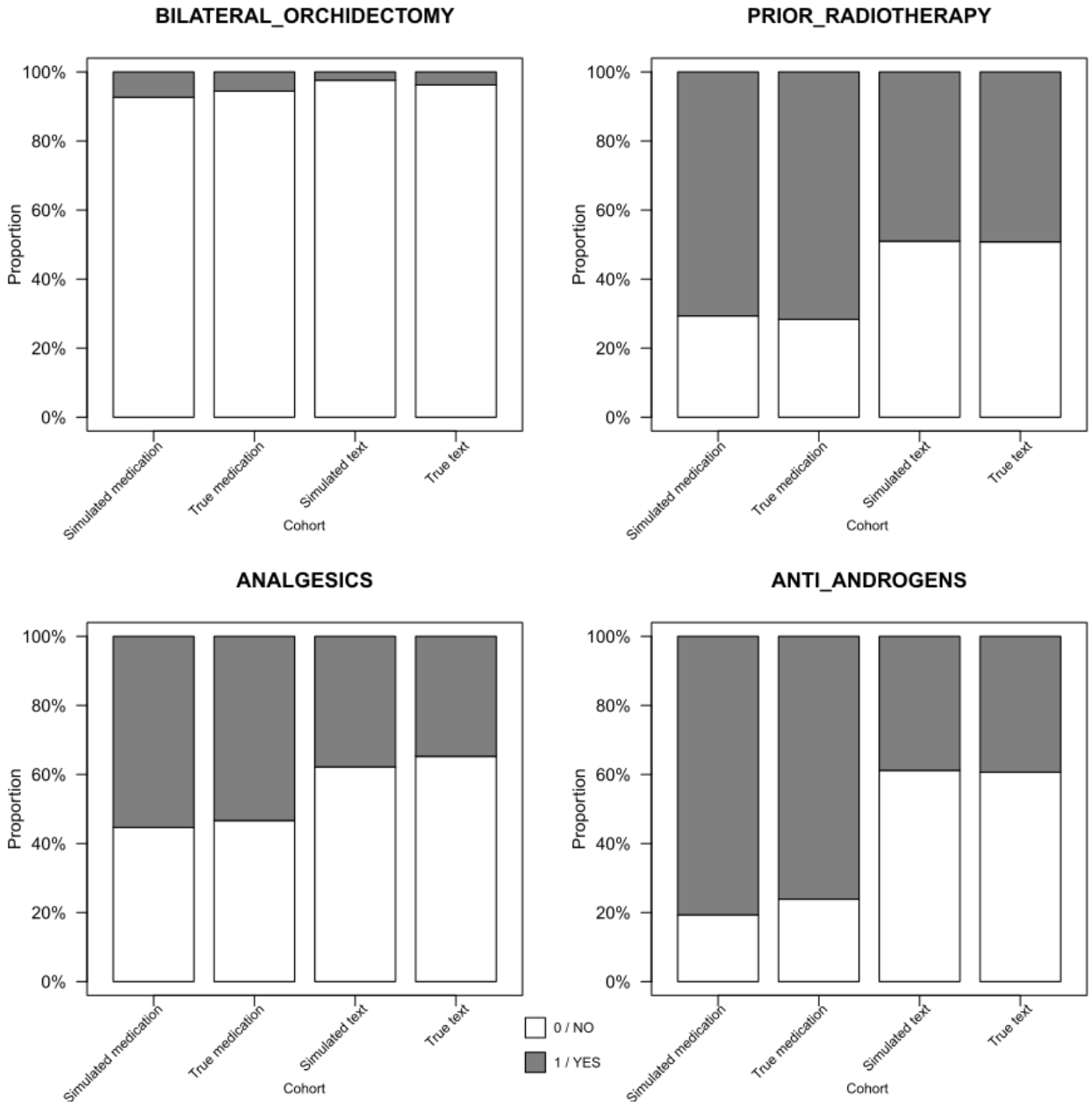
Supplementary Figure 2 (continued). Diagnostic boxplots for numeric variables in the simulated and true medication and text search-based cohorts.



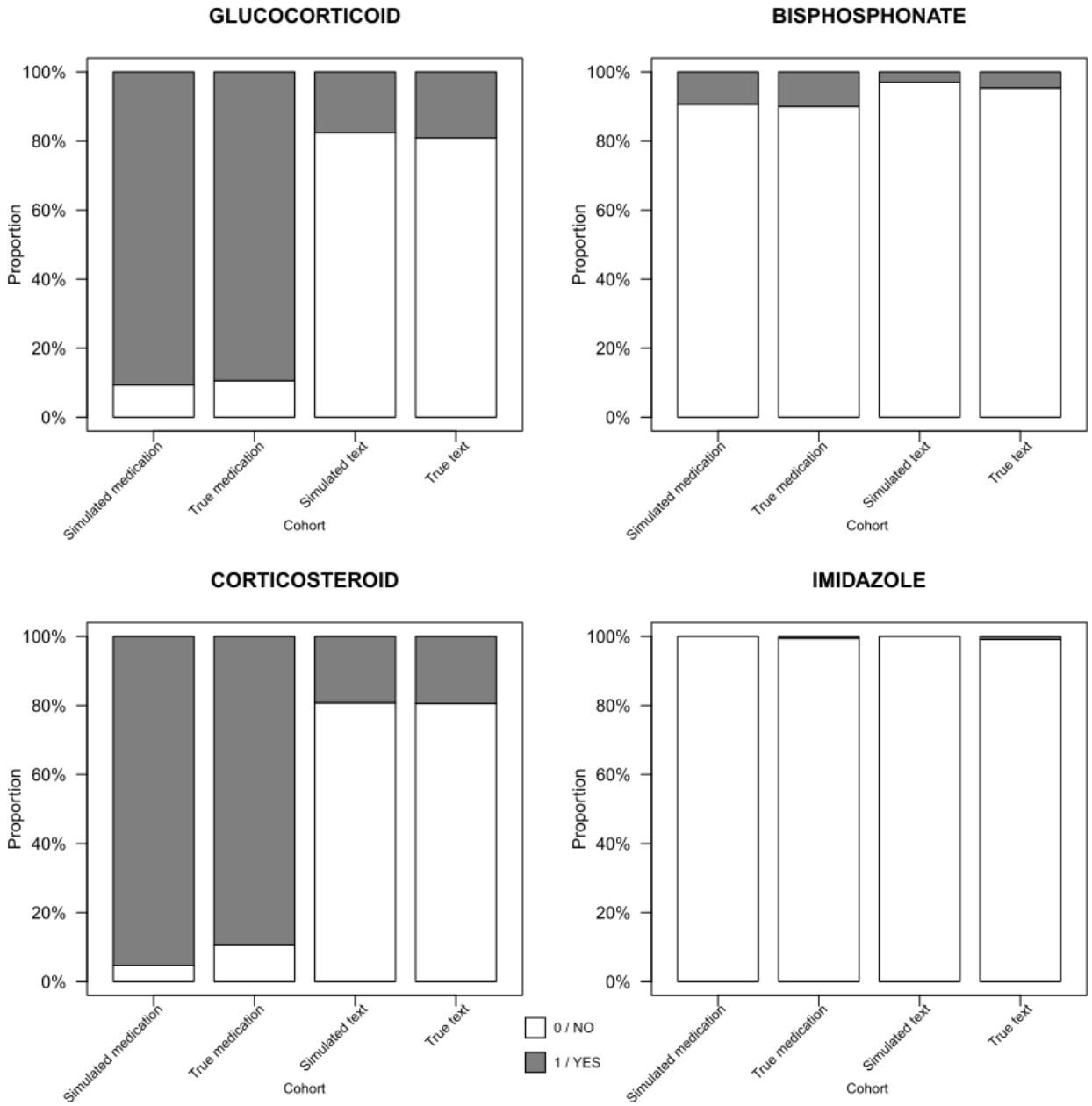
Supplementary Figure 3. Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



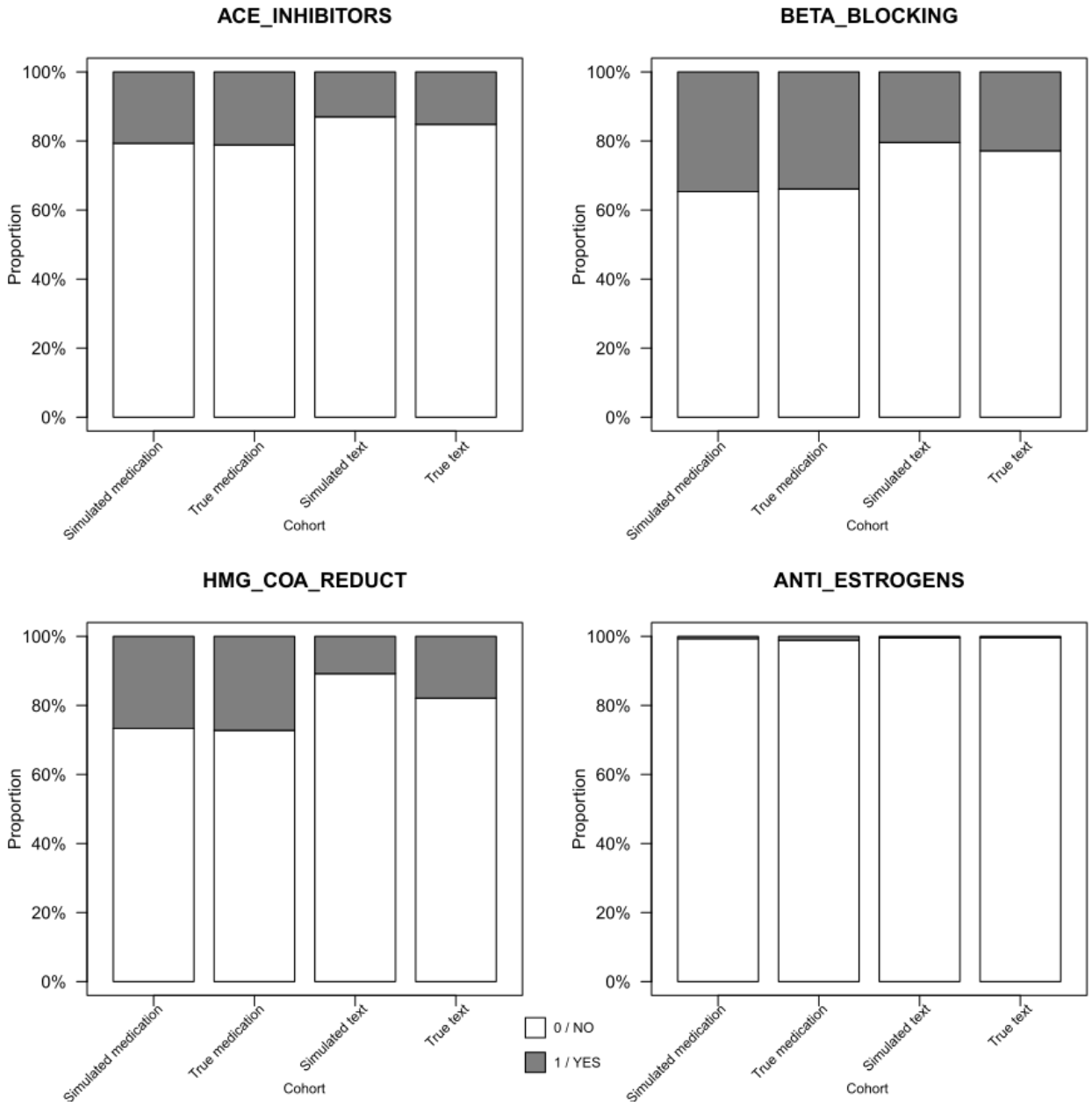
Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



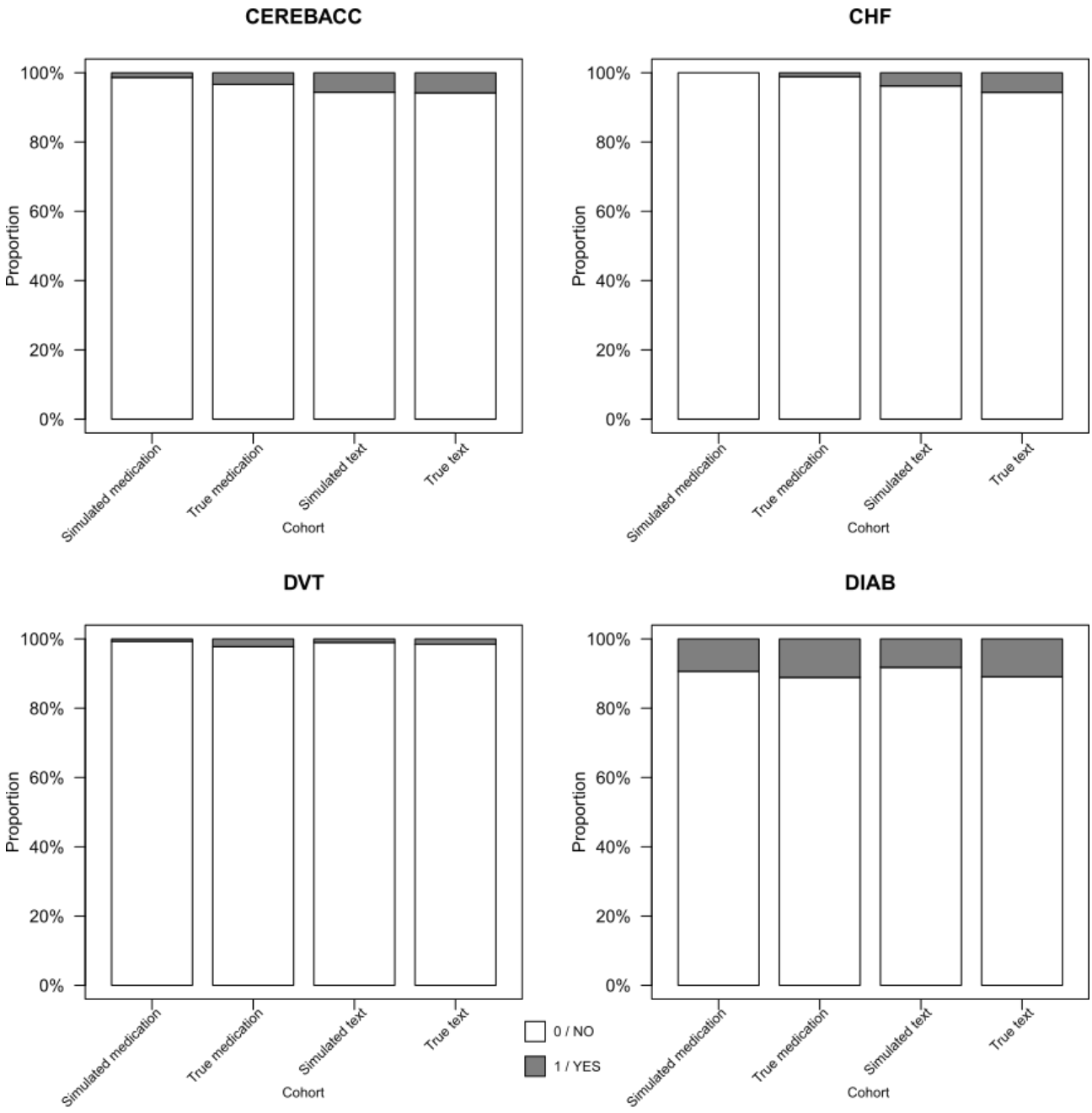
Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



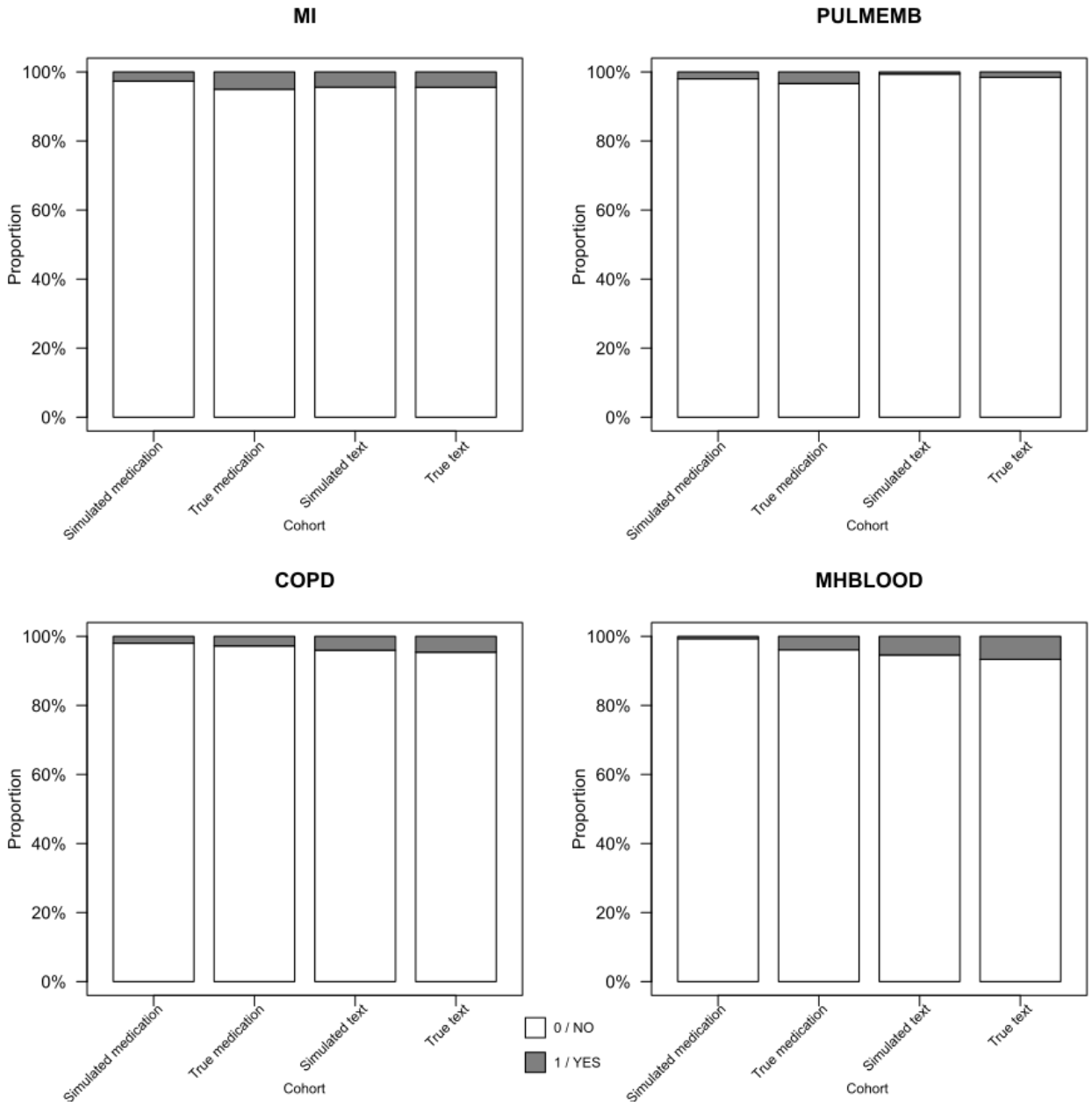
Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



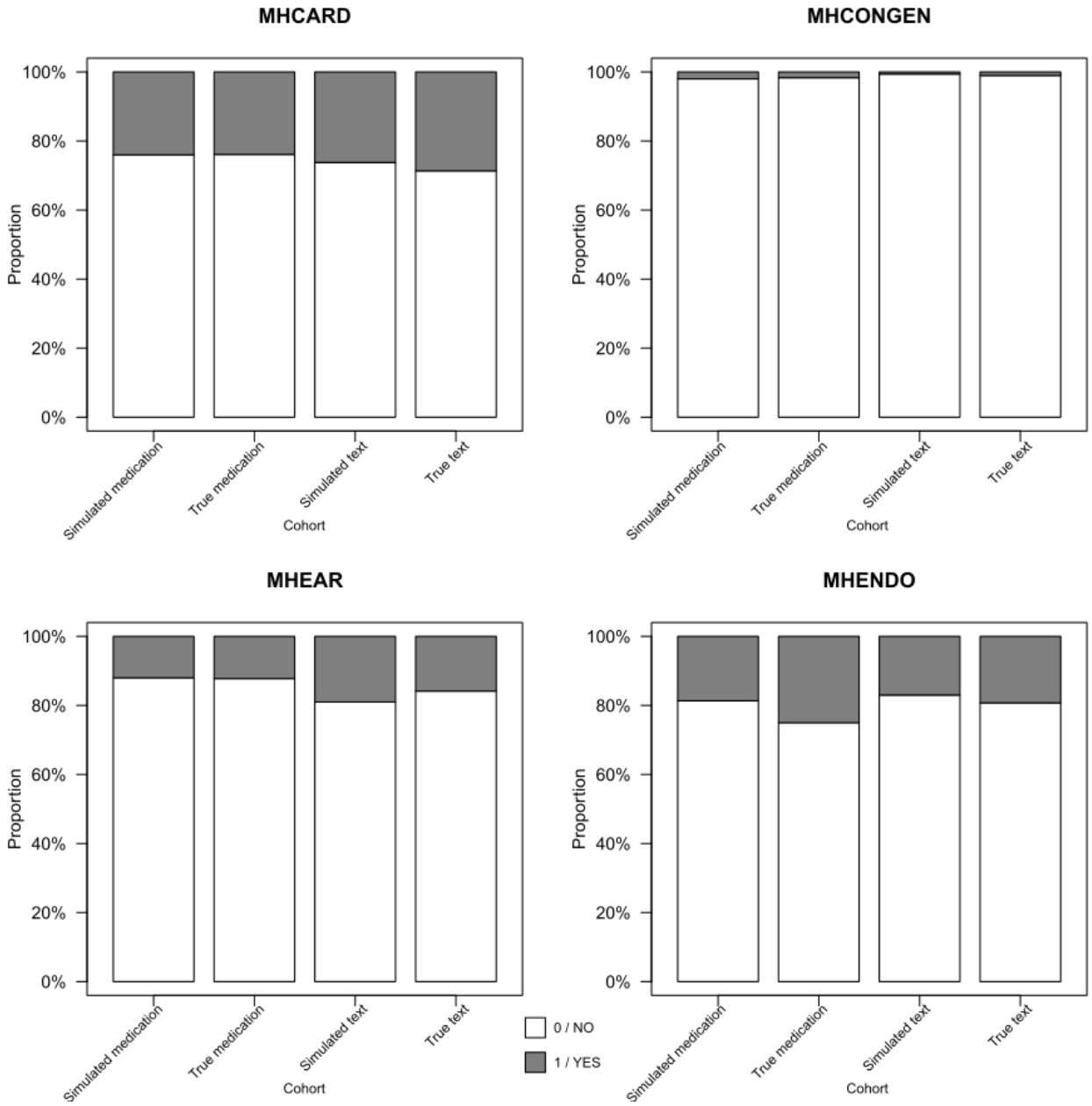
Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



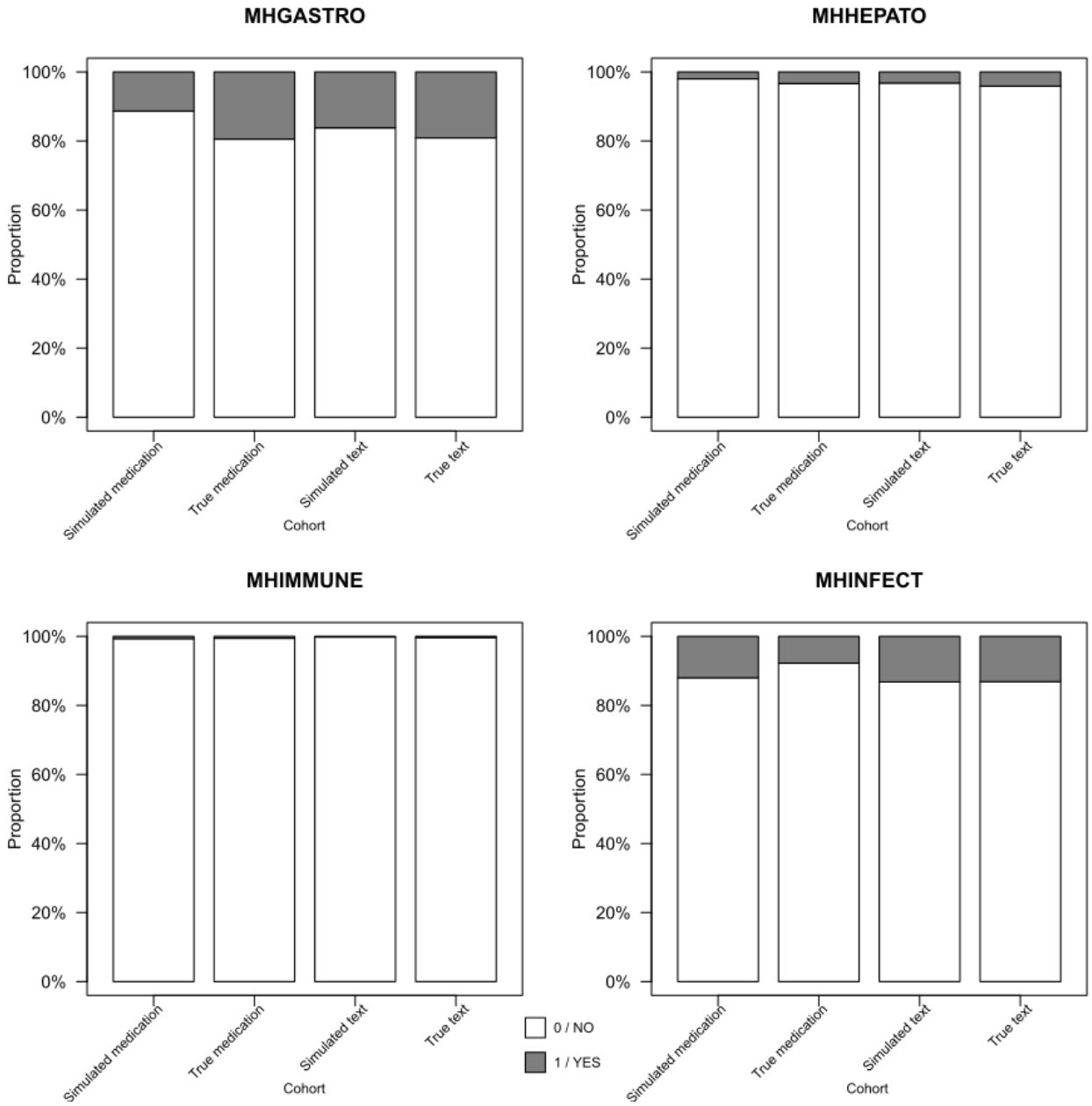
Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



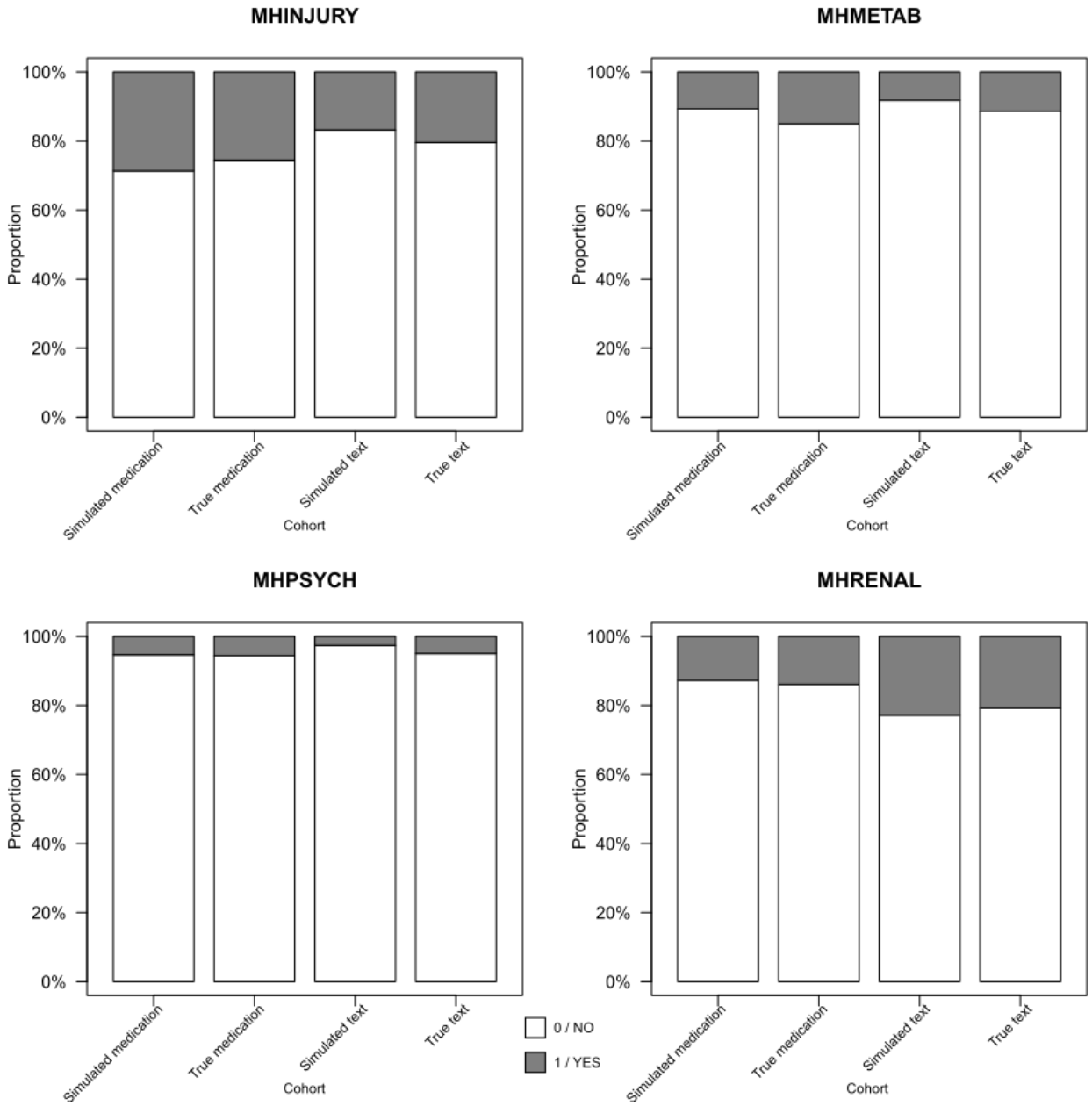
Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



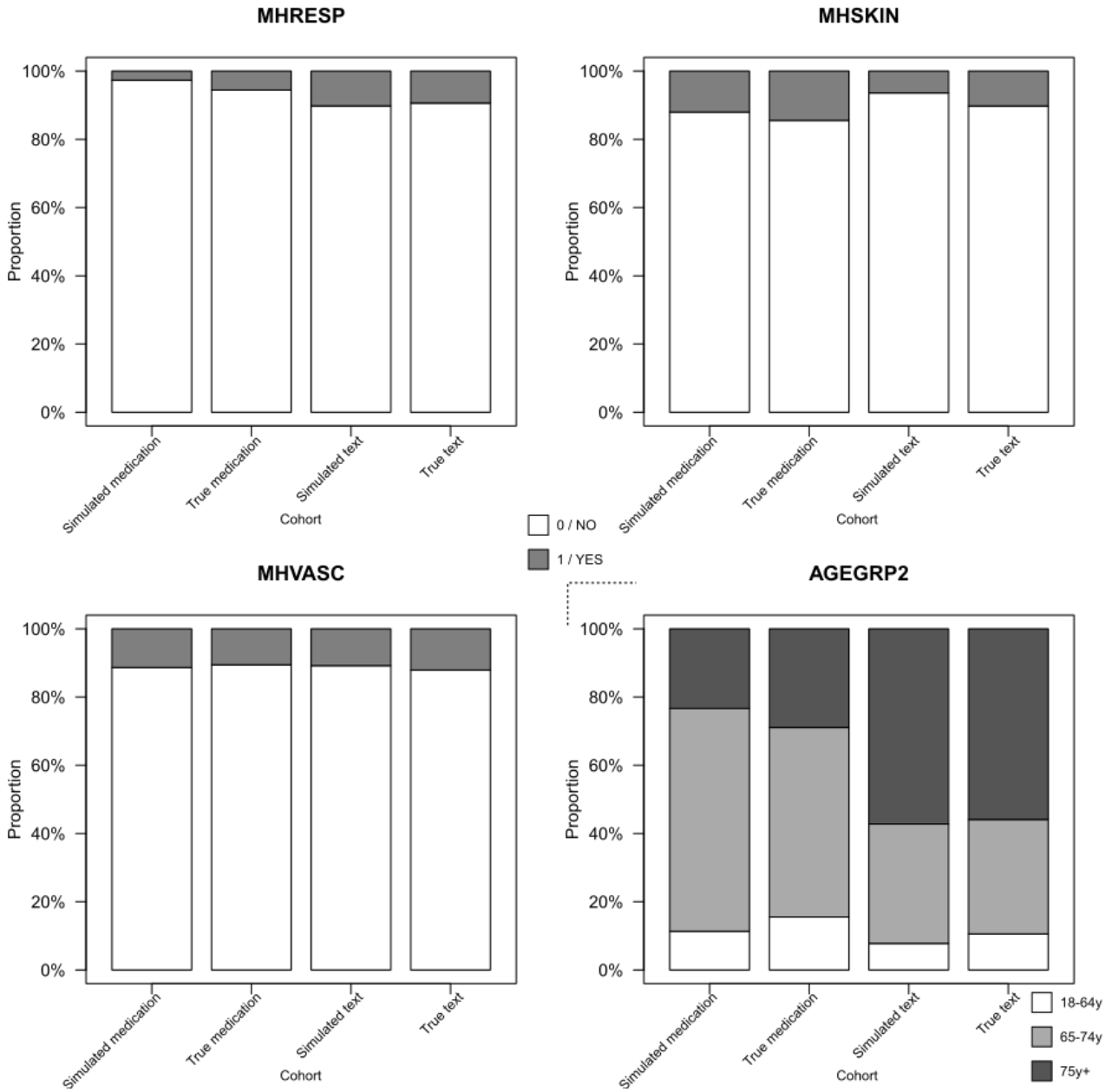
Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.



Supplementary Figure 3 (continued). Diagnostic bar-plots for the binary and ordinal features in the simulated and true medication and text search-based cohorts.