PLOS ONE

RESEARCH ARTICLE

Subtyping of common complex diseases and disorders by integrating heterogeneous data. Identifying clusters among women with lower urinary tract symptoms in the LURN study

Victor P. Andreev[1]*, Margaret E. Helmuth[1], Gang Liu[2], Abigail R. Smith[1], Robert M. Merion[1], Claire C. Yang[3], Anne P. Cameron[4], J. Eric Jelovsek[5], Cindy L. Amundsen[5], Brian T. Helfand[6], Catherine S. Bradley[7], John O. L. DeLancey[8], James W. Griffith[9], Alexander P. Glaser[6], Brenda W. Gillespie[10], J. Quentin Clemens[4], H. Henry Lai[11], The LURN Study Group[¶]

1 Arbor Research Collaborative for Health, Ann Arbor, Michigan, United States of America, **2** Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, United States of America, **3** Department of Urology, University of Washington, Seattle, Washington, United States of America, **4** Department of Urology, University of Michigan, Ann Arbor, Michigan, United States of America, **5** Department of Obstetrics and Gynecology, Duke University, Durham, North Carolina, United States of America, **6** Department of Urology, NorthShore University HealthSystem, Evanston, Illinois, United States of America, **7** Department of Obstetrics and Gynecology, University of Iowa, Iowa City, Iowa, United States of America, **8** Departments of Obstetrics and Gynecology and Urology, University of Michigan, Ann Arbor, Michigan, United States of America, **9** Department of Medical Social Sciences, Northwestern University, Chicago, Illinois, United States of America, **10** Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, United States of America, **11** Department of Surgery, Washington University, St Louis, Missouri, United States of America

¶ Membership of the LURN Study Group is provided in the Acknowledgments.
* victor.andreev@arborresearch.org

OPEN ACCESS

Citation: Andreev VP, Helmuth ME, Liu G, Smith AR, Merion RM, Yang CC, et al. (2022) Subtyping of common complex diseases and disorders by integrating heterogeneous data. Identifying clusters among women with lower urinary tract symptoms in the LURN study. PLoS ONE 17(6): e0268547. https://doi.org/10.1371/journal.pone.0268547

Editor: Peter F.W.M. Rosier, University Medical Center Utrecht, NETHERLANDS

Received: October 1, 2021

Accepted: May 3, 2022

Published: June 10, 2022

Data Availability Statement: The data that support the findings of this study are openly available in the NIDDK Central Repository at https://repository.niddk.nih.gov/; please reference the acronym "LURN".

Funding: This study is supported by the National Institute of Diabetes & Digestive & Kidney Diseases through cooperative agreements (grants DK097780, DK097772, DK097779, DK099932, DK100011, DK100017, DK099879) and for Dr.

Abstract

We present a methodology for subtyping of persons with a common clinical symptom complex by integrating heterogeneous continuous and categorical data. We illustrate it by clustering women with lower urinary tract symptoms (LUTS), who represent a heterogeneous cohort with overlapping symptoms and multifactorial etiology. Data collected in the Symptoms of Lower Urinary Tract Dysfunction Research Network (LURN), a multi-center observational study, included self-reported urinary and non-urinary symptoms, bladder diaries, and physical examination data for 545 women. Heterogeneity in these multidimensional data required thorough and non-trivial preprocessing, including scaling by controls and weighting to mitigate data redundancy, while the various data types (continuous and categorical) required novel methodology using a weighted Tanimoto indices approach. Data domains only available on a subset of the cohort were integrated using a semi-supervised clustering approach. Novel contrast criterion for determination of the optimal number of clusters in consensus clustering was introduced and compared with existing criteria. Distinctiveness of the clusters was confirmed by using multiple criteria for cluster quality, and by testing for significantly different variables in pairwise comparisons of the clusters. Cluster dynamics were explored by analyzing longitudinal data at 3- and 12-month follow-up. Five clusters of women with LUTS were identified using the developed methodology. None of the

Andreev's Biomarker Ancillary LURN R01 (grant 5R01DK125251). Research reported in this publication was supported at Northwestern University, in part, by the National Institutes of Health's National Center for Advancing Translational Sciences, grant UL1TR001422. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

clusters could be characterized by a single symptom, but rather by a distinct combination of symptoms with various levels of severity. Targeted proteomics of serum samples demonstrated that differentially abundant proteins and affected pathways are different across the clusters. The clinical relevance of the identified clusters is discussed and compared with the current conventional approaches to the evaluation of LUTS patients. The rationale and thought process are described for the selection of procedures for data preprocessing, clustering, and cluster evaluation. Suggestions are provided for minimum reporting requirements in publications utilizing clustering methodology with multiple heterogeneous data domains.

Introduction

Complex diseases, such as obesity, diabetes, atherosclerosis, Alzheimer's disease, major depressive disorder, and cancer result from multiple genetic, epigenetic, and environmental factors, and importantly, interactions between these factors [1–3]. Important differences between patients with these complex diseases and disorders may exist at multiple levels, including: (a) symptoms, subjective experiences, and adaptive behaviors; (b) characteristics of the physical state of the organism, including comorbidities; (c) characteristics of organs and systems; and (d) characteristics at the cellular or molecular level. The extent of these differences suggests that some complex diseases, disorders, and symptom complexes are better represented by subtypes, each of which can potentially have different etiologies, mechanisms, and outcomes, and require different approaches to treatment. Subtype identification is therefore a potentially important aspect to understanding and treating complex diseases and disorders. Specifically, extracting patient characteristics at each level and grouping patients based on these characteristics allows for comprehensive clinical phenotyping and provides necessary information for discovery and implementation of personalized treatments. Characteristics at each level are represented by variables of different types (continuous, categorical, and binary), scales, and different level of relevance or impact for the disease of interest. This data heterogeneity requires thoughtful preprocessing and novel approaches to data integration.

Lower urinary tract symptoms (LUTS) is a general term representing a heterogeneous group of symptoms or symptom complex with sometimes unclear etiology, high economic and social costs, and significant effects on patients' quality of life. LUTS can include frequent urination during day and night (nocturia), urinary urgency (a sudden urge to urinate), stress and urgency urinary incontinence (UI), and bladder emptying symptoms such as straining, hesitancy (delay to start to urinate), weak urine stream, and post-void dribbling. None of the symptoms is pathognomonic for a particular diagnosis, and many persons have more than one symptom. The prevalence of LUTS in the United States (US) ranges between 45% and 70% and increases with age [4,5]. Given the aging population in the US, the prevalence of LUTS is expected to increase in the coming years [6]. Medical expenditures for LUTS have been reported to be as high as $65 billion per year [7], yet many therapeutic options for the treatment of LUTS do not provide long-term symptom relief. Many patients experience a *combination* of symptoms, so treatment that focuses on a single symptom may result in suboptimal care. To improve treatment outcomes for patients with LUTS, it is necessary to sort out the heterogeneity of this population, increase the understanding of different subtypes of LUTS and their underlying mechanisms.

One way to better understand a complex disease, disorder, or symptom complex is to use an unbiased, data-driven unsupervised clustering approach to identify subtypes of individuals

with the disease. This method uses data to identify groups, or clusters, such that members of each cluster are as similar as possible to others within their cluster, but as different as possible from those in other clusters [8]. Subtypes identified in this manner are based on similarities and differences within the data, remaining agnostic to clinical definitions or diagnostic categories.

Clustering methodologies have generated important contributions to the analysis of health-related data and are becoming increasingly valuable tools as the field of precision medicine progresses. These methods represent a burgeoning field of research [9–15]. Many unsupervised classification or clustering methods have been developed, from commonly used k-means clustering, hierarchical clustering, and self-organizing maps (SOM) [16–18] to algorithms developed in specific areas for specific applications [19–21]. An important problem in unsupervised clustering is determining the optimal number of clusters. Currently available criteria include Calinski-Harabasz, Davies-Bouldin, Dunn, Gap, Silhouette indices, and others [22–26]; however, the optimal number of clusters may vary by the criterion applied. Therefore, another critical issue is to validate the clustering results, that is, to gain confidence about the clinical significance of the putative clusters, both in terms of cluster numbers and cluster assignments.

Disease subtyping by clustering "-omics" data of certain types, mostly gene expression data, has been extensively published [27–30]. More recent studies have subtyped complex diseases by clustering heterogeneous data that included patient health questionnaires and other clinical data. These research studies include subtyping asthma by using questionnaires, physiological tests, and lab tests [31]; subtyping type 2 diabetes by using body mass index (BMI), age at onset of diabetes, homoeostasis model assessment estimates, and insulin resistance [32]; and subtyping sepsis using demographics, vital signs, biomarkers of inflammation, and organ dysfunction or injury [33].

Resampling based consensus clustering initially proposed for clustering gene expression data [34] is gaining popularity and has been used for subtyping complex diseases using heterogeneous data [33,35]. Multiple random resampling of patients followed by k-means clustering generates probabilities that each pair of patients appear in the same cluster, which can be treated as a pairwise distance between patients and used for determination of cluster membership through hierarchical clustering [34]. This method ensures the clustering results are robust to sampling errors.

Previous studies aiming to identify subtypes of LUTS in an unbiased manner include Epidemiology Urinary Incontinence and Comorbidities (EPIC) and Boston Area Community Health (BACH) projects [36,37], which performed clustering of LUTS patients based on a relatively small number of self-reported symptom data in community-dwelling cohorts. Another study used only BMI and bladder diary variables for clustering community-dwelling women with LUTS [38].

The Symptoms of Lower Urinary Tract Dysfunction Research Network (LURN) Observational Cohort Study is a multi-center study that collected self-reported symptoms, 3-day bladder diaries, physical examination, neuroimaging and sensory testing data, and biological samples in over 1000 care-seeking men and women across six tertiary care centers. LURN is focused on defining patient-reported outcomes in people with lower urinary tract dysfunction (LUTD), conducting deep phenotyping of such individuals, and identifying biomarkers that are associated with symptoms of LUTD [39,40]. In our previous study [41], we performed clustering of 545 women from the LURN Observational Cohort Study using baseline self-reported urinary symptom data, captured with the LUTS Tool [42,43] and the American Urological Association Symptom Index (AUA-SI) [44]. Four distinct clusters were identified. Women in cluster F1 (n = 138) were continent, but reported post-void dribbling, frequency, and voiding

symptoms. Cluster F2 (n = 80) reported urgency urinary incontinence, as well as urinary urgency and frequency, and minimal voiding symptoms. Cluster F3 (n = 244) included women reporting all types of urinary incontinence, urgency, frequency, and mild voiding symptoms. Women in cluster F4 (n = 83) reported all LUTS at uniformly high levels. These subtypes of LUTS were based solely on the above two questionnaires and require further refinement, followed by clinical verification.

The current report describes the methodology and results of refining the female LUTS symptom-based clusters by integrating multiple data domains collected in LURN: demographics, non-urinary symptoms, history, and physical examination data, as well as intake and voiding patterns captured in 3-day bladder diaries. We discuss preprocessing of heterogeneous data and combination of continuous and categorical data using our novel weighted Tanimoto indices approach. We use resampling-based consensus clustering [34] combined with a modified semi-supervised clustering approach [45,46] to make use of data available on only a subset of participants. Then we determine the optimal number of clusters using our novel contrast criterion (CC) developed for consensus clustering, and compare it with other consensus clustering criteria: proportion of ambiguous clustering (PAC) [47], and consensus score (CS) [35], as well as with the established quality of clustering criteria, such as Calinski-Harabasz, Davies-Bouldin, Dunn, and Silhouette [22–26]. We identify distinct clusters of women with LUTS and show superiority of these clusters to our published symptom-based clusters [41], in terms of the percentage of significantly different variables in pairwise comparisons of the clusters and the confidence level in the determined cluster membership. Dynamics of the clusters in 12-month follow-up, as well as clinical relevance of the clusters, are discussed.

In the *Methods* section, we describe the analytical pipeline we developed and used for subtyping women with LUTS. We provide the rationale for our choices of methods for data preprocessing, integration, clustering, and cluster evaluation, as well as review of other available options. We demonstrate that the developed pipeline allowed for identification of distinct and robust refined clusters, with a higher percentage of significantly different variables across the clusters than those previously published, and validate cluster distinctiveness by analyzing biomarker data. Finally, we review the methodological information needed to assess subtyping via clustering and propose a set of reporting requirements that should ideally be included in all clustering reports. We finish by calling for the clustering community effort to develop minimum requirements for clustering publications. We believe this paper would be of interest for clinicians and researchers involved in subtyping of common complex diseases and disorders using heterogeneous and multidomain data.

Materials and methods

LURN data used for subtyping of LUTS

Data on women with LUTS. Data for LUTS subtyping were obtained from the LURN Observational Cohort Study [39,40], which included 545 women seeking care for LUTS at six tertiary care centers. Baseline data collection included demographic information, medical history, physical examination findings, 3-day bladder diaries [48], and self-report questionnaires of urologic and non-urologic symptoms. Urologic symptoms were collected using the LUTS Tool [42–43] and the AUA-SI [44]. The LUTS Tool contains 44 items, including questions on the frequency of occurrence and degree of bother for each urinary symptom. Possible answers to the LUTS Tool questions were ranked from zero to four, zero indicating absence of the symptom, and four indicating the most severe level of the symptom. The AUA-SI has eight items, including a single overall bother question. Responses to the first seven questions of the AUA-SI range from zero to five, zero indicating "none" or "not at all", and five indicating

"almost always". The final question in the AUA-SI ranges from zero (delighted) to six (terrible). Participants also completed patient-reported outcome (PRO) questionnaires from the Patient-Reported Outcomes Measurement Information System (PROMIS). We used questionnaires related to bowel function (PROMIS gastrointestinal constipation, diarrhea, and bowel incontinence subsets) [49], psychological health (PROMIS Depression and Anxiety Short Forms [50], Perceived Stress Scale [51], PROMIS Sleep Disturbance Short Form [52]), urologic pain (Genitourinary Pain Index [GUPI]) [53], and pelvic floor function (Pelvic Floor Distress Inventory [PFDI]) [54]. Demographics included: age, race, ethnicity, employment status, education level, and marital status. Physical examination data included: weight, waist circumference, BMI, post-void residual volume, pelvic organ prolapse (using the Pelvic Organ Prolapse Quantification [POP-Q] system that measures the location of selected landmarks on the vagina and cervix [55]), and presence of pathology findings at the introitus, urethra, vagina, uterus, or rectum. Medical history data included Functional Comorbidity Index (FCI) [56], additional individual comorbidities, as well as information on history of urinary tract infections (UTI), pregnancy, vaginal deliveries, alcohol, smoking, recreational drug use, and medication use. Individual comorbidities and medication use are categorial variables that were transferred into binary variables, e.g., comorbidity A (present or absent, 1 or 0), medication B (used or not used,1 or 0), and then clustered using the Tanimoto indices approach, described in the 'Clustering Pipeline' section below.

Bladder diaries included data on timing and volume of each beverage intake and urinary void during a 72-hour period. Completeness and accuracy of bladder diaries collected in LURN are described in [57]. Only 193 women (35%) returned bladder diaries deemed complete. For clustering purposes, we used the following five bladder diary variables from those 193 women: number of intakes and voids, total volumes of intakes and voids, and maximum voided volume (serving as a proxy for bladder capacity).

In total, 185 variables were used for subtyping women with LUTS; 27 demographic variables, 55 medical history variables, 33 physical exam variables, 52 urinary symptoms variables (LUTS Tool and AUA-SI), 13 non-urinary PRO variables, and 5 bladder diary variables. Variables were continuous (n = 83) or categorical (n = 102). S1 Table in S1 File presents an overview of these variables for 545 women with LUTS used in the analysis. Although all variables were deemed important or possibly important in subtyping persons with LUTS, not all the variables are of equal importance, relevance, and non-redundancy; therefore, scaling and weighting of the variables was implemented as described below.

Data on non-LUTS controls. Our preprocessing procedure, described in more detail in the next section, includes scaling of variables by standardizing their values using means and standard deviations (SDs) in non-LUTS controls. The LURN study included 55 control women, who were not necessarily healthy but did not report LUTS. Unfortunately, not all the variables of interest were collected for these non-LUTS controls (e.g., physical examination, bladder diary). As a source of bladder diary data for non-LUTS controls, we used bladder diaries of 32 non-LUTS controls from the Establishing Prevalence of Incontinence (EPI) community study of women in Southeastern Michigan [58]. For other variables of interest not collected for non-LUTS controls, we used population data from literature sources indicated in S1 Table in S1 File.

Biological samples collected and analyzed in LUTS cases and non-LUTS controls

The LURN study collected numerous biological samples, including whole blood, serum, plasma, and urine at baseline and at 3- and 12-month follow-up visits [39,40]. Of these

samples, 230 baseline serum samples of women with LUTS and 30 serum samples of non-LUTS controls were analyzed using the targeted proteomics approach–Proximity Extended Assay (PEA) by Olink Proteomics (Uppsala, Sweden). Three Olink panels (cardio metabolic, inflammation, neurology) were used to quantify abundances of 276 proteins. These data were not used for clustering in the current report; however, they served as an additional orthogonal approach for evaluation of the quality of the identified clusters. We compared the abundances of 276 proteins in women with LUTS and in controls and tested for significantly differentially abundant proteins in each of the identified clusters versus controls adjusted for multiple comparison using the false discovery rate (FDR) correction (FDR<0.05) [59]. Note that assays were performed in a subset (n = 230, 42%) of women.

Ethical guidelines and consent

The authors confirm all relevant ethical guidelines have been followed, and all research has been conducted according to the principles expressed in the Declaration of Helsinki. Informed written consent has been obtained from participants. Institutional Review Board (IRB) approval has been obtained from: Ethical and Independent Review Services (E&I) IRB, an Association for the Accreditation of Human Research Protection Programs (AAHRPP) Accredited Board, Registration #IRB 00007807.

Overview of the clustering pipeline

The clustering pipeline implemented in this paper contains multiple steps of data preprocessing, integration, clustering, and cluster evaluation. In this subsection, we present an overview of the sequence of steps in the pipeline shown in Fig 1. Details and rationale for each of the steps are provided in the rest of the subsections of the "Materials and Methods" section below. The pipeline is implemented using MATLAB (MathWorks, Natick, MA) and is publicly available through the Dryad repository (https://datadryad.org).

Data preprocessing

Multiple imputation. Due to missing data (up to 10% in self-report questionnaires of urologic and non-urologic symptoms), multiple imputation was performed. The imputation used a sequential regression technique and was implemented using IVEware version 2.0 [60,61]. Ten imputed data sets were constructed, and each was preprocessed as described below. The k-means step of the resampling-based consensus clustering was performed on each of the data sets separately, resulting in ten pairwise probabilities of being in the same cluster for each pair of participants. Finally, hierarchical clustering on the mean of ten probabilities (pairwise distances) was performed to determine cluster membership.

Scaling of continuous variables. Scaling variables prior to clustering is an important and often overlooked step. Clustering algorithms group objects in a way that minimizes the sum of the pairwise distances between participants within the cluster, where the distance is composed of the distances between all variables calculated using Euclidian, Manhattan, or other suitable metrics [9]. Since each variable is measured using its own scale, the distances and optimal partition of the objects depend on these scales. As stated in [62], the problem with unscaled, unstandardized data is the inconsistency between cluster solutions when the scale of some of the variables is changed, which is a strong argument in favor of standardization. It is especially important in the case of heterogeneous data, where scales of variables in the raw data can be very different and completely unrelated. A common form of conversion of variables to standard scores (or z-scores) entails subtracting the mean and dividing by the SD for each variable. However, subtracting the cohort mean and dividing by the cohort SD would mask the subtype

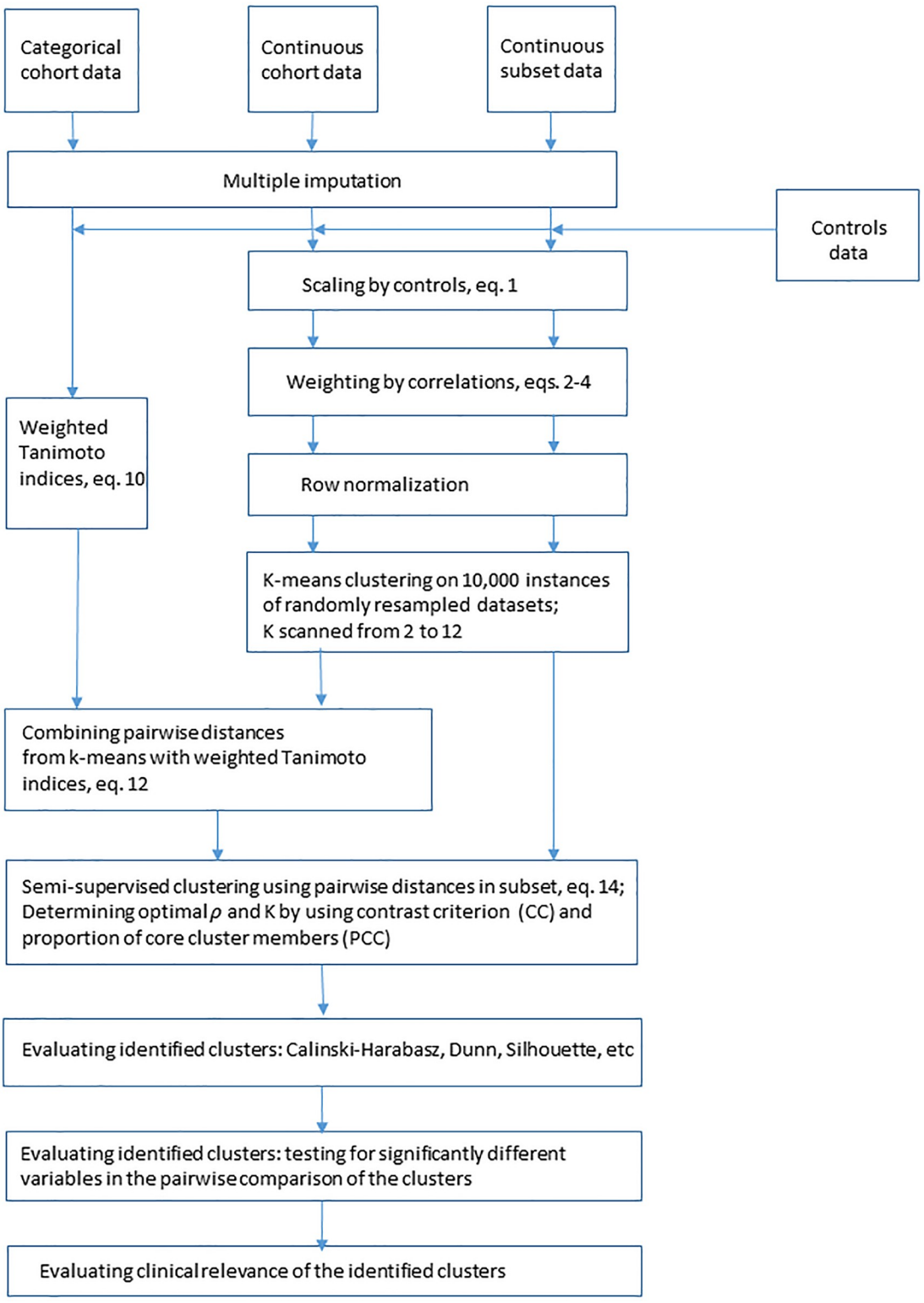


Fig 1. Flowchart of the pipeline for subtyping of a common complex disease or disorder by integrating heterogeneous data as used for subtyping of LUTS. Three types of data are imputed. Continuous variables are scaled using controls, weighted, normalized, and then clustered using consensus k-means clustering. Categorical data is transformed into binary and then clustered using weighted Tanimoto indices approach. Matrices of pairwise distances for three types of data are then integrated to maximize contrast criterion (CC) and proportion of core cluster members (PCC). Identified clusters are evaluated using several clustering criteria and testing for significantly different variables in the pairwise comparison of the clusters.

https://doi.org/10.1371/journal.pone.0268547.g001

differences, since it ignores whether the within-cohort variance is caused by the natural biological variability of the subjects or by differences in disease subtypes, which increase the within-cohort variance due to multimodal distributions along certain variables. Using z-scores along such variables will unduly reduce their weight and will mask the presence of the subtypes. Therefore, standardization using z-scores is not suitable for our task of identifying disease subtypes. The solution to this problem is standardization by the mean and SD of a reference population that does not have the disease of interest, in this case, controls without LUTS. Following this approach, we define standardized variables $\boldsymbol{S_{in}}$ [63]:

$$S_{in} = (A_{in} - A_{iC})/\sigma_{iC} \tag{1}$$

where A_{in}–i^{th} unstandardized variable for participant n, A_{iC}-mean value of i^{th} variable across control subjects without disease of interest, σ_{iC}–SD of i^{th} variable in control subjects without disease of interest. A simulated example illustrating the benefits (substantially lower misclassification error) of clustering using variables standardized according to equation (Eq 1), versus unstandardized variables and z-scores, is presented in Supplemental Material text in S1 File.

Weighting variables to mitigate the redundancy. Clustering results can be skewed by including variables reflecting redundant information. An obvious and extreme case example will be including into the data set the same or highly correlated variables multiple times, which will result in the dominating role of these variables in the overall sum of squared distances, and therefore in the clustering decision. To mitigate this, we used weighting, so that the highest weight was attributed to the least correlated variable (i.e., the variable with the smallest average correlation with all other variables) and the lowest weight to the most correlated variable [41,64]. The weights w_i were defined by Eqs (2–4):

$$w_i = \frac{1}{1 + c_i/c_w} \tag{2}$$

$$c_i = \sum_{j=1, j\neq i}^{m} r_{ij}/(m-1) \tag{3}$$

$$c_w = \sum_{i=1}^{m} \sum_{j=1, j\neq i}^{m} r_{ij}/m(m-1) \tag{4}$$

where m -number of variables and r_{ij} Pearson correlation coefficients of variables i and j.

Row normalization for continuous variables. Initial attempts to cluster un-normalized urinary symptoms data led to identification of two clusters that differed by overall severity of LUTS, not subtypes of symptoms. To avoid clustering predominantly by the overall severity of LUTS, in [41,64] and here, we normalized the data by the participant's overall severity of disease. For each participant, the weighted Euclidean length of the vector composed of all 78 continuous variables used for clustering was calculated as $\boldsymbol{L_n} = \sqrt{\sum_{i=1}^{78} (\boldsymbol{w_i} \cdot \boldsymbol{S_{in}})^2}$, where $\boldsymbol{S_{in}}$ is the scaled i^{th} variable for the n^{th} participant and $\boldsymbol{w_i}$ is the weight of i^{th} variable defined by Eqs (2–4). Each continuous variable was then normalized by $\boldsymbol{L_n}$, the Euclidean length of the participant's vector, resulting in normalized continuous variables $\boldsymbol{V_{in}} = \boldsymbol{w_i} \cdot \frac{S_{in}}{L_n}$. This normalization strategy allowed for clustering based on the direction rather than the length of the vector representing each subject.

Consensus clustering using continuous variables

Clustering was performed using a resampling-based consensus clustering method introduced by Monti et al [34]. We performed 1000 instances of random resampling, each selecting a subset including 80% of participants. The same procedure was repeated for each of the ten multiply imputed data sets, resulting in 10,000 subsets. We then partitioned each of the subsets into clusters using a k-means clustering algorithm implemented as *k-means* MATLAB function (with option 'number of replicates' = 8, see [63] for the explanation on the need of this option); with number of clusters K scanned from 2 to 12. Let Q_{nq} denote the number of times participants n and q were assigned by k-means into the same cluster. Let I_{nq} denote the number of times participants n and q were both selected in the random sampling. The probability of participants n and q belonging to the same cluster could be calculated as Q_{nq}/I_{nq}. We could thus obtain ten probabilities for a pair of participants from the ten imputed data sets. The average of these probabilities represented the final consensus index M_{nq} for participants n and q. A 545 by 545 consensus matrix M (Fig 2) was constructed to visualize these average probabilities as a heat map. Probability is color-coded: bright yellow represents probability close to one and dark blue probability close to zero. The indices of participants were reordered so that the participants belonging to the same clusters were grouped together. To reorder the indices of participants in the consensus matrix, we employed hierarchical clustering (using *clustergram* MATLAB function) with $1-M$ as distance matrix so that participants belonging to the same clusters were grouped together, depicted as bright yellow blocks along the diagonal of consensus matrix.

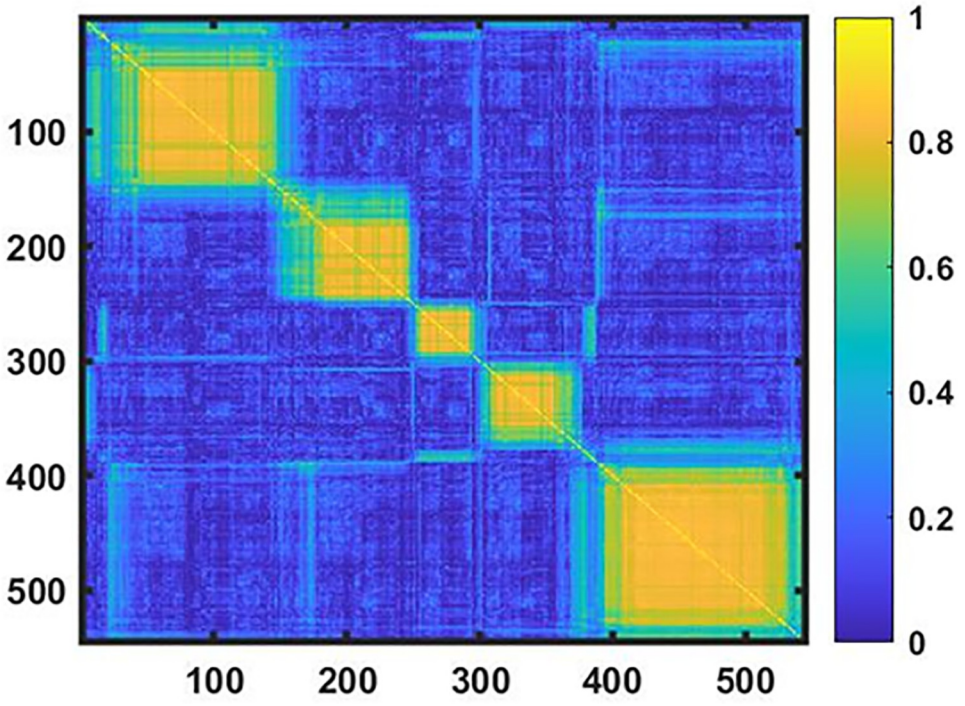


Fig 2. Consensus matrix. Consensus (545x 545) matrix is presented as a heat map, where the probabilities M_{nq} for each pair of participants to be in the same cluster are shown by color-coded elements; bright yellow represents probability close to one and dark blue probability close to zero. Five yellow squares along the diagonal represent 5 clusters of participants with LUTS.

https://doi.org/10.1371/journal.pone.0268547.g002

Clustering using categorical variables

Of 185 variables used for clustering women with LUTS, 102 (55%) are categorical. K-means is not an appropriate method for categorical variables, so resampling-based consensus clustering with multiple runs of k-means algorithm, which we used for continuous variables, cannot be directly used for categorical variables.

k-*prototype approach.* One way to combine continuous and categorical variables is to use the k-prototype algorithm introduced by Z. Huang [65]. According to [65], the distance between two objects ***X***, ***Y***, described by p continuous variables and m-p binary variables, is represented as:

$$d(X,Y) = \sum_{j=1}^{p}(x_j - y_j)^2 + \gamma\sum_{j=p+1}^{m}\delta(x_j, y_j) \tag{5}$$

where δ is Kronnecker symbol describing simple matching and γ is the weight introduced in [65] to "avoid favoring either type of attribute" (continuous vs. categorical). The goal of the k-prototype algorithm is to minimize the sum of the distances (defined by Eq 5) between objects within the cluster. Limitations of the k-prototype algorithm include lack of scaling of categorical variables and use of the same weight γ for all categorical variables, regardless of their relevance to the disease of interest or their redundancy. A later version of the algorithm [66] attempted to overcome some of these limitations by defining the distance between an object X_n and the centroid Z_k of the k^{th} cluster as:

$$d(Z_k, X_n) = \sum_{j=1}^{p}(x_{nj} - Z_{kj})^2 + \sum_{j=p+1}^{m}\varphi(Z_{kj}, x_{nj}) \tag{6}$$

where $\varphi\left(Z_{kj}, x_{nj}\right) = \begin{cases} 1, if\ Z_{kj} \neq x_{nj} \\ 1 - \dfrac{C_{kjr}}{C_k}, otherwise \end{cases}$, and C_k is the number of objects in cluster k, while C_{kjr} is the number of objects in this cluster with the categorical value a_j^r of the j^{th} attribute, e.g., participants with blue eyes in cluster k. Such definition of distance makes sense; for instance, if the number of participants in cluster k is 100, and 51 of them have blue eyes, then for a person with brown eyes, distance from the centroid along this dimension is 1, but for a person with blue eyes, it is 1–0.51 = 0.49. Now, if 99 participants have blue eyes, then for them, the distance from the centroid is 1–0.99 = 0.01, while for the only one with brown eyes, it is still 1. Thus, such a definition of distance makes certain attributes more important (defining) for the cluster if the majority of objects have the same value of this attribute. An algorithm using such a definition of distance between objects would strive to make clusters as homogeneous as possible with regard to both its continuous and categorical variables. This approach, however, does not distinguish between categorical variables relevant and irrelevant to the disease of interest.

To distinguish between relevant and irrelevant variables, we suggest using the same approach as for continuous variables, i.e., compare them with controls without the disease or symptom complex of interest. For the categorical variables transformed into binary variables, we suggest scaling by comparison of the frequencies of these binary variables in LUTS F_j and in controls F_{jC} by using the following function:

$$\gamma_j = \left|\log\left(\frac{F_j}{F_{jC}}\right)\right| \tag{7}$$

where $|x|$ is absolute value of x.

If, for a certain binary variable, frequencies in LUTS and non-LUTS controls are equal, e.g., prevalence of blue eyes is the same in LUTS and non-LUTS, then this variable will get weight

$\gamma_j = 0$ and would not affect clustering decision, essentially excluding the variable from clustering. However, if the frequency of this variable in LUTS is higher or lower than in controls, then $\gamma_j > 0$, and this, relevant to disease variable, will affect clustering decisions. To accommodate this scaling together with weighting of the variables based on their correlation with other variables described by Eqs (2–4), one needs to modify Eq (6):

$$d(Z_k, X_i) = \sum_{j=1}^{p} w_j (x_{ij} - Z_{jk})^2 + \sum_{j=p+1}^{m} w_j \varphi(Z_{kj}, x_{ij}) \gamma_j \tag{8}$$

Unfortunately, none of the available implementations of k-prototype algorithm in standard software (R and SAS [67,68]) easily allows for such modification, and therefore, an alternative simpler approach was used.

Weighted Tanimoto indices approach. A simpler approach to clustering categorical variables is based on Tanimoto indices or Tanimoto similarity measure [69]. For two objects a and b described by m binary variables, Tanimoto similarity is defined as:

$$T = \frac{\sum_{j=1}^{m} a_j \cdot b_j}{\sum_{j=1}^{m} (a_j^2 + b_j^2 - a_j \cdot b_j)} \tag{9}$$

For instance, if a and b are 5-dimensional binary vectors a = [1, 1, 0, 0, 0] and b = [1, 0, 1, 0, 0], then $T = 1/(2+2-1) = 1/3$. Note that common "ones" but not common "zeros" are counted in this definition of similarity, which is especially useful in case of multiple binary variables formed from one categorical variable. Think, for example, of the categorical variable 'eye color' transformed into several binary variables: 'blue eyes' (yes, no), 'brown eyes' (yes, no), 'green eyes' (yes, no), etc. Tanimoto similarity between two persons with blue eyes will not depend on whether you add 'hazel eyes' to the list of options or not.

Not all of the categorical variables are equally relevant to the disease of interest, so we want to be able to assign weights reflecting the level of relevance for each variable by comparing its frequency in LUTS with its frequency in non-LUTS controls. We also want to compensate for redundancy in the variables by using weights defined by Eqs (2–4). Importantly, we want to make sure that no categorical variable, even if it is much more frequent in disease than in controls, has overwhelmingly high weight and makes the role of differences in other variables negligible. To attain this goal, we introduce a weighed Tanimoto similarity measure as:

$$T = \frac{\sum_{j=1}^{m} a_j \cdot b_j \cdot w_j^2 \cdot (\mathrm{erf}(\gamma_j))^2}{\sum_{j=1}^{m} ((a_j^2 + b_j^2 - a_j \cdot b_j) \cdot w_j^2 \cdot (\mathrm{erf}(\gamma_j))^2)}, \tag{10}$$

where w_j is the weight defined by Eqs (2–4), using the appropriate correlation coefficients of the variables. Coefficient γ_j is defined by Eq (7) and minimizes the role of binary variables equally prevalent in disease and controls. Function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$ ensures that the weight of each binary variable is smaller or equal to one, even for the high values of γ_j, since $|\mathrm{erf}(x)| \leq 1$. Note that maximum value of T is equal to one when all the categorical variables in a and b are the same, and minimum value is zero when all the categorical variables are different. Now we can use Eq (10) to define distance between any pair of participants described by J binary variables as:

$$d\left(X_n, X_q\right) = 1 - T_{nq} = 1 - \frac{\sum_{j=1}^{J} X_{jn} \cdot X_{jq} \cdot w_j^2 \cdot (\mathrm{erf}(\gamma_j))^2}{\sum_{j=1}^{J} ((X_{jn}^2 + X_{jq}^2 - X_{jn} \cdot X_{jq}) \cdot w_j^2 \cdot (\mathrm{erf}(\gamma_j))^2)} \tag{11}$$

Combining continuous and categorical variables

Note the similarity of the pairwise distance $1-T_{nq}$ between two participants based on the categorical variables describing them (Eq 11), and pairwise distance $1-M_{nq}$ between these participants based on their continuous variables. The former is equal to zero when all the categorical variables describing these two participants are the same, while the latter is equal to zero when the two participants always were assigned to the same cluster by the 10,000 instances of k-means in the resampling-based consensus clustering. Similarly, the former is equal to one when all the categorical variables describing the two participants are different, while latter is equal to one when these participants always were assigned to the different clusters in resampling-based consensus clustering. This similarity allows combining continuous and categorical pairwise distances into a single distance measure D_{nq}, with a minimum value of zero and maximum value of one using the weighted Euclidean length approach:

$$D_{nq} = \sqrt{((1 - M_{nq})^2 + \mu^2 \cdot (1 - T_{nq})^2)/(1 + \mu^2)}, \quad (12)$$

where μ is the weight representing the relative role of the distances based on categorical variables and on continuous variables. We set it equal to the ratio of the number of non-redundant categorical and continuous variables: $\mu = \sum_{j=1}^{102} w_{jcat} / \sum_{j=1}^{78} w_{jcont}$, where w_j are determined by Eqs (2–4) using correlation coefficients appropriate to the distributions of the variables.

Semi-supervised clustering using bladder diary data

Of 545 women in the Observational Cohort of LURN, 193 (35%) returned complete bladder diaries without missing volumes of voids and intakes. These data were deemed clinically important to the subtyping of women with LUTS. One way to integrate data domains only available on a subset is by using semi-supervised clustering methods [70,71]. If class membership for the members of the subset is known, then the objective function in the clustering of the whole cohort should be modified as follows:

$$WCSS = \sum_{j=1}^{p} \sum_{n=1}^{Nk} \sum_{q=1}^{Nk} (x_{nj} - x_{qj})^2 + \sum_{n=1}^{Nsk} \sum_{q=1}^{Nsk} h_{nq}, \quad (13)$$

where *WCSS*—within cluster sum of squared distances, *p*–number of variables, $n \neq q$, *Nk*—number of cohort participants in the given cluster k, *Nsk*—number of the participants in cluster k that were present in the subset. The values of $h_{nq} = -h$ are negative (reward, decreasing the within-cluster-sum-of-squares [*WCSS*]) if participants *n* and m belong to the same cluster according to subset classification, and is positive $h_{nq} = h$ (punishment, increasing *WCSS*) if participants *n* and *q* belong to the different classes of the subset. This approach known as "must-link, cannot-link" allows for using labels known from classification of the subset to influence clustering of the cohort. The limitation of this approach, however, is that it does not allow for different level of confidence in subset cluster membership, i.e., participants *n* and *q* are either in the same subset cluster or not. In our case, additional data available for the subset of participants do not provide 100% confidence in cluster membership for this subset; furthermore, the number of participants in the subset is lower than in the whole cohort, making the subset clusters less robust. There is a measure of similarity, however, that is quantitative and reflects the confidence in cluster membership; it is the pairwise distance between members of the subset. We suggest using this measure to modify the pairwise distance defined by Eq 12 by taking into account the similarity between members of the bladder diary (BD) subset:

$$G_{nq} = \max\left((D_{nq} + \rho \cdot (BD_{nq} - mBD)), 0\right) \quad (14)$$

where D_{nq} is defined by Eq 12, BD_{nq} is the pairwise distance between members of the subset, *mBD* is the mean pairwise distance between members of the subset, ρ is a parameter determined as described in the below sections. If either participant *n* or *q*, or both of them are not members of the subset, their pairwise distance is not known and is assumed equal to the mean pairwise distance within the subset *mBD*. For these participants, the second term of Eq 14 is equal to zero; while, for members of the subset, it is either negative or positive depending on whether BD_{nq} is smaller or larger than *mBD*. Note that, since the second term might be negative, for some large values of ρ, the sum of the two terms is negative as well. However, the pairwise distance between the objects cannot be negative, and is therefore set to zero for these cases.

We used five variables (number of intakes and voids, total volumes of intakes and voids, and maximum voided volume) from the bladder diaries of the 193 participants to refine the values of their pairwise distances. These five variables were scaled using bladder diary data for controls [58] and then added to the 78 other continuous variables to calculate pairwise distances BD_{nq} refined with bladder diary variables, as described in the subsection on consensus clustering using continuous variables. Then it was introduced into Eq 14 to get the refined matrix of the pairwise distances G_{nq}. The value of the coefficient ρ was determined by optimizing the quality of clusters, as defined in the following subsections.

Determining the number of clusters

Determining the number of clusters is an important step in any clustering process. In partitioning algorithms like k-means, it is necessary to decide on the number of clusters *K* prior to running the algorithm. In agglomerative algorithms like hierarchical clustering, it is possible to decide on the number of clusters when the dendrogram based on the distances between objects is already created. It is common to try several values of *K* and then to compare the resultant clusters by using various quality of clustering criteria, including Calinski-Harabasz, Davies-Bouldin, Dunn, Gap, and Silhouette indices [22–26]. Quite often, these criteria disagree on the value of *K* that optimizes the quality of the clusters.

Resampling-based consensus clustering, introduced in [34] and subsequently applied to our data set, is a combination of multiple instances of k-means clustering followed by hierarchical clustering on the pairwise distances between objects derived at the first stage. The value of *K* in k-means is typically scanned (in our case from 2 to 12), and then the optimal value of *K* is determined using criteria developed specifically for consensus clustering algorithm [34,35,47]. In both [34] and [47], the determination of the number of clusters is based on analysis of the cumulative distribution function (CDF) of consensus index values M_{nq} (defined in the "Consensus clustering using continuous variables" subsection of this paper). The main idea of this analysis is that, in case of ideal clustering, there are only two possible values of consensus index $M_{nq} = 1$, when a pair of objects *n* and *q* are in the same cluster, and $M_{nq} = 0$, when they are in the different clusters. Therefore, the ideal CDF should consist of two vertical lines and a horizontal (flat) line between them. The length of the first vertical line will be equal to the number of pairs with $M_{nq} = 0$ and the length of the second vertical line to the number of pairs with $M_{nq} = 1$. However, if the value of *K* used in clustering is different from the true value of *K*, then the shape of the CDF curve differs from the idealized curve described above. In [34], the optimal *K* is defined as the one for which the change in the area under the CDF (relative to area at *K-1* and *K+1*) is the largest ("elbow" of the AUC vs. K curve). Analysis in [47] demonstrates several examples when the criterion of [34] does not work and suggests an alternative criterion named proportion of ambiguously clustered pairs (PAC) equal to the number of pairs with $0<M_{nq}<1$ over the total number of pairs. Obviously, in the real-world

situation of biological variability and noisy measurements, almost all of the pairs will fall into this category, so some more liberal lower and upper boundaries for M_{nq} need to be introduced, e.g., 0.1 and 0.9, as in [47]. It raises a question, however, whether these boundaries should be different for different values of K, since potential ambiguity increases with the increased number of clusters.

A more straightforward approach to determine the number of clusters is used in [35] by introducing mean consensus score (CS) calculated as the mean value of consensus indices M_{nq} within the clusters. The value of K that results in the highest CS is considered optimal. The problem with CS is that it favors the high number of clusters, e.g., CS could reach its maximum when K = N/2 and each cluster contains only a pair of most similar objects with highest values of M_{nq}.

Below, we introduce a pair of complementary criteria, i.e., contrast criterion (CC) and proportion of the core clusters (PCCs), which we used in our clustering pipeline. We believe they combine the advantages of PAC and CS and are free of some of their limitations, especially when used as a complementary pair. The below sections introduce the CC and PCCs based on the analysis of consensus matrix derived by k-means clustering of the multiple resampling instances of the data set; however, these criteria can be used with any matrix of pairwise distances between the objects.

Contrast criterion. The idea of the CC is derived from visual representation of the consensus matrix as a heat map presented in Fig 2. Each pixel of the heat map represents the value of $\boldsymbol{M_{nq}}$ probability of two objects to be together in the same cluster. Each bright yellow square along the diagonal of the matrix represents a cluster. Next, we compared the "bright yellowness" of this diagonal square with the "color" of the rest of the row in which the diagonal square is located; therefore, the term "contrast criterion" (CC). The larger the difference between these two measures, the further the situation from the case where all $\boldsymbol{M_{nq}}$ except for n = q are equal, and the heat map is uniformly yellowish (single uniform cluster case), in which case the contrast is zero. We consider number of clusters K and cluster membership optimal when CC is maximized.

When analyzing the properties and behavior of clustering criteria, it is necessary to compare the clusters identified using certain clustering algorithms and clustering criteria of interest with the "true" clusters. Unfortunately, "true" clusters are not known in real life, and therefore, one needs to simulate them and then evaluate misclassification error resulting from the clustering algorithm and criteria of interest. Such an approach was used in [63] to compare several popular clustering algorithms, and was applied for the case of resampling-based consensus clustering with CC and PAC criteria (see Supplemental Material text and S2-S10 Figs in S1 File). Here, we concentrate on the general analysis of CC and its properties. In this analysis, we need to introduce the term "alleged clusters", which are different from "true clusters" and "identified clusters". "Alleged clusters" are determined for each value of K tried by clustering algorithm, while "identified clusters" are those maximizing the value of clustering criteria of interest, and "true clusters" are specified by the simulation.

To define CC explicitly, let us first look at the most typical case where the number of alleged clusters is not equal to the number of objects, is not equal to one, and none of the clusters includes just one object. Note that we are not making any assumptions or imposing any restrictions on the properties of the "true" clusters.

$$\text{For } K \neq N, K \neq 1, N_j \neq 1,\ CC = \sum_{k=1}^{K} \left\{ \frac{\sum_{n=1}^{N_k} \sum_{q=1, q\neq n}^{N_k} M_{nq}}{N_k(N_k-1)} - \frac{\sum_{i=1, i\neq k}^{K} \sum_{n=1}^{N_k} \sum_{q=1}^{N_i} M_{nq}}{N_k(N-N_k)} \right\} / K, \quad (15)$$

where K is the number of alleged clusters, N is the number of the clustered objects, N_k and N_i

are the numbers of objects in k^{th} and i^{th} clusters. As intended, the first term in the brackets of Eq (15) represents the average "bright yellowness" of the diagonal square, while the second term in the brackets represents the average "yellowness" of the rest of the row in which the diagonal square is located.

Note that we defined CC as an averaged contrast across alleged K clusters independently of the size of the clusters. Other approaches are possible, e.g., a weighted average based on the sizes of the clusters, or minimax approach, where the contrast for the "worst" (least contrast cluster) is maximized. Clearly, the best choice of combining contrasts of each of the clusters into one overall depends on many factors, including the goal of the clustering, expected sizes of the clusters, and the number and distribution of the variables. It is an interesting topic; however, it is outside of the scope of this paper.

Another choice made in defining CC by Eq (15) is omitting of the diagonal terms M_{nn}, which are always equal to one. The percentage of diagonal terms in each square representing an identified cluster is $\frac{N_k}{N_k^2} = N_k^{-1}$, or 100%, 50%, 33%, 25% for $N_k = 1, 2, 3, 4$. Since $M_{nn} = 1$, $M_{nq}<1$, inclusion of diagonal terms in the definition of CC would favor smaller clusters over larger clusters by assigning higher CC to the smaller clusters, irrespective of similarities of objects within the clusters. Therefore, we do not include diagonal terms in the definition of CC, as shown in Eq 15. Note that Eq (15) does not work if $K = 1$, $K = N$, or $N_j = 1$. For these special cases, CC is defined and discussed in Supplemental Material text in S1 File.

There are certain similarities between our contrast criterion CC and consensus score (CS) of [35]. Although the exact definition of CS is not provided, it appears that CS is similar to the first term of Eq (15). However, it is unclear if the diagonal terms M_{nn} are omitted or included and how CS for K clusters are combined to derive the overall CS. Nevertheless, it is of interest to compare the behavior of CS and CC in some simple idealized cases. In case of ideal clustering, both CS and CC reach their maximum possible value of 1. The worst-case scenario for both criteria is the case where $K>1$ clusters are alleged, when in reality, there are no "true" clusters, and all objects are described by a unimodal random distribution of their attributes (variables). Now, if the number of alleged clusters $K>1$, there is an equal probability for the object to end up in any of K clusters, so the mean value of M_{nq} is $1/K$, which makes the minimal possible value for CC equal to zero and minimal possible value for CS equal to $1/K$. The rather limited range of values from $1/K$ to 1, together with the dependence of the minimal value on the number of alleged clusters, are the limitations of the CS criterion, which are absent in case of CC, where the range is from 0 to 1 irrespectively of K. Therefore, we consider the use of CC advantageous for determining the optimal number of clusters.

Core clusters. When analyzing the quality of alleged clusters, it is important to know the confidence with which cluster membership is determined. Clearly, one would prefer clusters where the probability of objects to belong to a particular cluster is 0.9 rather than 0.3. The knowledge of consensus matrix allows for calculating the probability for each object n belonging to a particular cluster k:

$$\pi_{nk} = \frac{\sum_{q=1 q\neq n}^{N_k} M_{nq}}{N_k - 1} \tag{16}$$

Note that it is different from the probability averaged across the cluster that was used to calculate CS and CC in the previous subsection. Importantly, within the same cluster, some objects might have probability (confidence) as high as $\pi_{nk} = 0.9999\ldots$, or as low as $\pi_{nk} = 1/K + 0.001$, assuming that it is lower for any other cluster $i\neq k$. We will call the n^{th} object the core member of cluster k if $\pi_{nk}>0.5$, which means that, for this object, the probability to be in cluster k is higher than probability to be in all other clusters combined. The rest of the members of

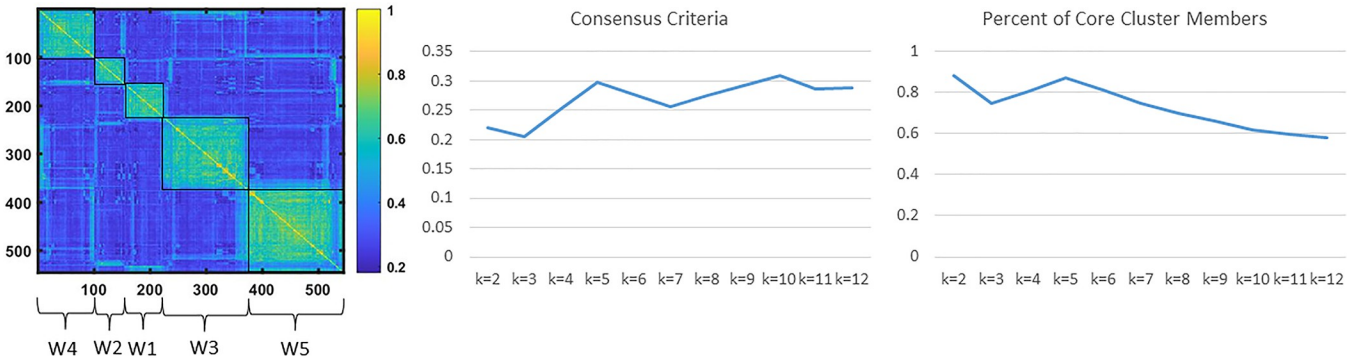


Fig 3. Determination of the optimal number of refined clusters. (A) Consensus matrix heat map demonstrates five clusters of participants (named W1-W5) grouped together based on the pairwise distances G_{nq} (Eq 14). (B) Contrast criterion (CC Eq 15) for K = 2,. . .,12. (C) Proportion of core cluster members (PCC Eq 16) for K = 2,. . .12. Both CC and PCC have maxima at K = 5, justifying the selection of five clusters.

https://doi.org/10.1371/journal.pone.0268547.g003

the cluster do not belong to the core; for them, the probability to belong to the k^{th} cluster is just higher than to be in any other given cluster. The number of core members divided by the total number of objects in the cluster provides a useful measure that we named proportion of the core cluster (PCC). As with the contrast criterion CC, one can use several approaches to derive overall PCC from the PCCs for each cluster, i.e., take the average across all clusters, weighted average based on the size of the clusters, or minimax by looking at the PCC in the worst cluster. PCC provides a measure of overall confidence in the alleged cluster membership and of the uniformity of the alleged clusters. Unlike contrast CC, PCC favors smaller size of the clusters and reaches its maximum when K = N/2 and each cluster contains just a pair of objects. PCC reveals information similar to the proportion of ambiguously clustered pairs (PAC) [47], for which $0.1<M_{nq}<0.9$; however, PCC is easier to interpret and does not include unjustified upper and lower boundaries 0.1 and 0.9.

Using CC and PCC to determine optimal number of clusters K. We used a combination of CC and PCC to determine the optimal number of clusters K. Note that Eq 14 contains undefined coefficient $\boldsymbol{\rho}$. Therefore, we have two parameters $\boldsymbol{\rho}$ and K to determine and two criteria to meet. We determined $\boldsymbol{\rho}$ and K as the values maximizing CC and corresponding to an elbow (point of diminishing returns) for PCC. A clustering procedure was run for 24 values of $\boldsymbol{\rho}$ from 0.05 to 1.2 using the single-program multiple data sets (spmd) function of MATLAB Parallel Computing Toolbox; K values were scanned from K = 2 to K = 12. The determined optimal values were $\boldsymbol{\rho}$ = 0.3 and K = 5. The resultant consensus matrix together with the values of CC and PCC for K scanned from 2 to 12 are presented in Fig 3. As seen in Fig 3, contrast criterion CC and percent of core cluster members PCC both have maxima for number of clusters K = 5. As shown in Table 1, other quality of clustering criteria also confirm K = 5 as an optimal number of clusters for our cohort of women with LUTS.

Table 1. Other quality of clustering criteria, confirming K = 5 as an optimal number of clusters.

	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7	K = 8	K = 9	K = 10
Calinski-Harabasz (↑)	**1382.26**	346.13	676.29	**1126.25**	567.83	398.02	309.93	234.52	180.57
Davies-Bouldin (↓)	**0.4328**	0.8175	0.6045	**0.4720**	0.5692	0.9178	1.0809	1.1498	1.1678
Dunn (↑)	0.0167	0.0072	0.0290	**0.0485**	**0.0496**	0.0455	0.0273	**0.0754**	**0.0590**
Point-Biserial (↓)	**-4.801**	-2.728	-3.25	**-3.849**	-2.515	-2.081	-1.855	-1.621	-1.409
Silhouette (↑)	**0.6896**	0.4565	0.5731	**0.6857**	0.5955	0.4821	0.3723	0.3143	NaN

https://doi.org/10.1371/journal.pone.0268547.t001

Evaluation of the quality of the identified clusters

We used multistep procedure to evaluate the quality of identified clusters and compare with the potential alternative clusters. The first step was examination of CC and PCC for each of K clusters (Fig 3). Next, we calculated other established quality of clustering criteria, including Calinski, Davies-Bouldin, Dunn, Point-Biserial, Silhouette [22–26], etc. (Table 1). Then, we performed pairwise comparison of the clusters using Wilcoxon rank sum tests and chi square tests, where appropriate, to determine which variables used for clustering were significantly different in the pairwise comparison. All pairwise comparisons were adjusted using an FDR correction for multi-testing [59].

Visualization of the results

We used several tools to visualize the results of our analyses. Heat maps representing the consensus matrices were generated using the clustergram MATLAB function. Properties of the identified clusters were illustrated by radar plots, built-in using SAS statistical graphics panel (sgpanel) scatter, polygon, and vector procedures. Comparison of cluster membership in the refined clusters versus previously identified symptom-based clusters was performed using Sankey diagrams, built with the googleVis package.

Results and discussion

Description of the clusters

Five distinct clusters of women with LUTS were identified by clustering 545 participants of the LURN Observational Cohort Study using 185 variables. We call these clusters W1-W5, in order to distinguish them from clusters F1-F4, described previously [41]. Demographic data for each cluster are presented in Table 2. Some demographic characteristics were different across the clusters, including age, ethnicity, menopausal status, obesity, prevalence of hysterectomy, percentage of participants with at least one vaginal birth, education level, and employment status. No significant differences across the clusters were observed for race, smoking, and alcohol use.

Importantly, all urinary symptoms and many other clinical variables were significantly different across the clusters. Table 3 presents the comparison of urinary symptoms (collected with LUTS Tool and AUA-SI), bladder diary variables, and other 37 significantly different clinical variables across clusters W1-W5. For clarity, we describe these significantly different variables while discussing signatures of the clusters in the text following Table 3. Table 2, and especially Table 3, illustrate the distinctiveness of the identified five clusters of women with LUTS.

Properties of the five clusters are visualized in Fig 4. Each column represents one of five clusters. Radar plots in the first row illustrate urinary symptoms measured by LUTS Tool and AUA-SI; the second row illustrates demographics, clinical measurements, and non-urinary PROs; the third row shows categorical data on comorbidities and anomalies identified during the physical exam; the fourth row shows intake and voiding pattern variables collected in bladder diaries. Radar plots represent mean values of the raw variables across members of each of the clusters. None of the clusters could be characterized by a single symptom, but rather by a combination of symptoms with various levels of severity. Women in all five clusters reported higher than normal frequency of voiding (with the highest frequency in W3 and W5). Women in all clusters except W1 reported urinary urgency and some level of incontinence.

Women in cluster W1 (n = 77) reported minimal urinary incontinence, but had mostly voiding and post-micturition symptoms (post-void dribbling, trickling, straining, hesitancy,

Table 2. Demographic data for clusters W1-W5.

	W1	W2	W3	W4	W5	P-Value
N	77	64	144	95	165	
Age (median IQR)	53 (36–60)	66 (51–71)	60 (50–70)	51 (42–63)	59 (51–67)	<0.0001
Race						0.221
White	66 (86%)	60 (94%)	116 (81%)	80 (84%)	130 (79%)	
Black	8 (10%)	2 (3%)	22 (15%)	8 (8%)	26 (16%)	
Asian	3 (4%)	2 (3%)	3 (2%)	5 (5%)	3 (2%)	
Other	1 (1%)	2 (3%)	3 (2%)	3 (3%)	6 (4%)	
Ethnicity						0.016
Non-Hispanic/Latino	74 (96%)	57 (89%)	138 (96%)	84 (88%)	159 (96%)	
Hispanic or Latino	1 (1%)	5 (8%)	3 (2%)	10 (11%)	2 (1%)	
Unknown	2 (3%)	2 (3%)	3 (2%)	1 (1%)	4 (2%)	
Obese	23 (30%)	15 (23%)	72 (50%)	38 (40%)	99 (60%)	<0.0001
Post-menopausal	35 (46%)	46 (73%)	102 (71%)	47 (49%)	117 (72%)	<0.0001
Had a hysterectomy	17 (22%)	9 (14%)	49 (34%)	25 (26%)	64 (39%)	0.0013
At least one vaginal birth	35 (45%)	60 (94%)	91 (63%)	75 (79%)	128 (78%)	<0.0001
Alcoholic drinks per week						0.0706
Never	10 (13%)	5(7%)	20 (14%)	15 (17%)	40 (25%)	
0–3 drinks per week	55 (72%)	46 (73%)	89 (62%)	65 (68%)	100 (61%)	
4–7 drinks per week	8 (11%)	9 (14%)	21 (15%)	13 (14%)	15 (9%)	
More than 7 drinks per week	0 (0%)	3 (4%)	8 (6%)	2 (2%)	5 (3%)	
Smoking status						0.1746
Never smoker	50 (67%)	43 (68%)	99 (69%)	64 (67%)	92 (56%)	
Former smoker	21 (28%)	19 (30%)	33 (23%)	28 (29%)	52 (32%)	
Current smoker	3 (4%)	1 (2%)	10 (7%)	3 (3%)	18 (11%)	
Education						0.0058
Less than Associate degree	40 (30%)	24 (30%)	89 (30%)	27 (28%)	76 (47%)	
Associates or Bachelor's degree	32 (43%)	25 (40%)	58 (41%)	45 (47%)	61 (38%)	
Graduate degree	23 (31%)	19 (30%)	41 (29%)	23 (24%)	24 (15%)	
Employment						0.0350
Full-time	35 (47%)	27 (42%)	50 (35%)	44 (46%)	51 (31%)	
Part-time	13 (17%)	5 (8%)	20 (14%)	17 (18%)	21 (13%)	
Unemployed (looking or not looking for work)	27 (36%)	32 (50%)	73 (51%)	34 (36%)	90 (56%)	

https://doi.org/10.1371/journal.pone.0268547.t002

and incomplete bladder emptying). They were younger than the average across the LURN female cohort, had a lower than average weight, number of pregnancies, and vaginal births. They had less comorbidities and abnormal findings in the physical exam. Women in cluster W2 (n = 64) reported mild urinary symptoms, including mild urinary incontinence. They presented clinically significant anterior vaginal wall prolapse (mean POP-Q point B anterior [Ba] = 1.24 cm, which is outside the introitus), apical prolapse (mean POP-Q point C = -2.38 cm), and the most severe pelvic organ prolapse symptoms (with the highest Pelvic Organ Prolapse Distress Inventory [POPDI-6] values of 20.60). They were on average older (66 vs. 53 years old), and had a higher number of pregnancies (2.9 vs. 1.8) and vaginal births (1.47 vs. 0.92) than women in cluster W1. They also had the highest post-void residual urine volume (75 mL) across the clusters. Women in cluster W3 (n = 144) reported high urinary frequency, urinary urgency, and urgency urinary incontinence. They had increased weight, had larger waist circumference, and higher functional comorbidity index (FCI) than women in W1, W2, and W4.

Table 3. Urinary symptoms, bladder diary variables, non-urinary symptoms, and clinical variables across clusters W1-W5.

	Cluster W1	Cluster W2	Cluster W3	Cluster W4	Cluster W5	P-Value
N	77	64	144	95	165	
Frequency	1.92	1.72	2.65	1.67	2.70	<0.001
Daytime frequency	1.67	1.44	2.09	1.37	2.13	<0.001
Nocturia	1.85	1.16	1.89	1.22	2.06	<0.001
Incomplete emptying	1.76	1.16	1.20	0.86	2.37	<0.001
Trickle/dribble	1.69	1.07	1.28	1.11	2.56	<0.001
Urgency	1.10	1.47	2.88	1.76	2.73	<0.001
Hesitancy	1.18	0.72	0.34	0.27	1.26	<0.001
Intermittency	1.12	0.73	0.32	0.28	1.28	<0.001
Straining	1.04	0.49	0.16	0.11	1.09	<0.001
Weak stream	0.84	0.79	0.43	0.22	1.52	<0.001
Spraying	0.87	0.60	0.46	0.45	1.55	<0.001
Urgency with fear of leakage	0.56	1.34	2.85	2.05	2.69	<0.001
Bladder pain	0.94	0.50	0.34	0.34	1.42	<0.001
Burning with urination	0.44	0.24	0.10	0.12	0.72	<0.001
Urinary incontinence (UI)	0.59	1.24	2.39	2.81	2.96	<0.001
Post-void UI	0.74	0.65	0.71	1.44	2.31	<0.001
Urgency UI	0.37	1.06	2.61	2.00	2.71	<0.001
Stress UI (laughter)	0.55	1.03	1.26	2.86	2.57	<0.001
Stress UI (exercise)	0.46	0.78	0.90	2.94	2.35	<0.001
UI with sleep	0.10	0.35	0.73	0.88	1.64	<0.001
UI with sex	0.16	0.33	0.12	0.75	0.80	<0.001
UI no reason	0.21	0.56	0.89	1.66	2.13	<0.001
Nocturia (AUA-SI)	2.37	1.77	2.18	1.57	2.37	<0.001
Frequency (AUA-SI)	3.15	2.25	2.96	2.26	3.13	<0.001
Intermittency (AUA-SI)	1.74	1.12	0.53	0.41	1.95	<0.001
Weak stream (AUA-SI)	1.25	1.12	0.60	0.36	2.09	<0.001
Straining (AUA-SI)	1.44	0.57	0.12	0.16	1.08	<0.001
Incomplete emptying (AUA-SI)	2.18	1.28	1.04	0.79	2.67	<0.001
Urgency (AUA-SI)	1.59	2.17	3.63	2.18	3.53	<0.001
QOL (AUA-SI)	3.68	3.57	4.46	4.66	4.88	<0.001
Weight (Kg)	75.58	75.47	82.62	80.58	87.64	<0.001
Waist circumference (Cm)	94.74	97.13	100.59	95.58	106.35	<0.001
Systolic blood pressure	122.42	131.00	130.92	124.29	130.17	<0.001
Post-void residual volume (mL)	55.43	75.04	41.29	29.09	39.94	0.012
POP-Q: Ba result	-2.65	1.24	-2.23	-1.78	-1.86	<0.001
POP-Q C result	-6.85	-2.38	-6.52	-6.54	-5.85	<0.001
POP-Q D result	-7.96	-4.43	-7.79	-7.42	-6.79	<0.001
Number of pregnancies	1.80	2.90	2.01	2.48	3.07	<0.001
Number of vaginal births	0.92	2.43	1.47	1.72	2.02	<0.001
Functional Comorbidity Index Total	1.66	1.74	2.48	1.24	3.64	<0.001
GUPI pain	5.09	4.15	3.70	2.23	6.84	<0.001
GUPI urine	4.71	3.40	3.67	2.49	5.67	<0.001
GUPI QOL	5.78	5.68	6.85	7.60	8.70	<0.001
POPDI-6	14.50	20.60	7.57	7.25	30.22	<0.001
Colorectal-anal distress inventory (CRADI-8)	12.98	18.03	14.76	10.99	33.50	<0.001
Urinary distress inventory (UDI-6)	24.42	25.44	36.27	39.58	63.67	<0.001

(*Continued*)

Table 3. (Continued)

	Cluster W1	Cluster W2	Cluster W3	Cluster W4	Cluster W5	P-Value
Perceived stress scale	12.09	9.29	11.89	11.29	16.44	<0.001
PROMIS constipation T-score	50.30	48.68	49.00	49.89	55.80	<0.001
PROMIS depression T-score	48.91	44.63	48.67	46.61	53.93	<0.001
PROMIS anxiety T-score	49.84	46.78	48.61	48.11	54.75	<0.001
PROMIS sleep disturbance T-score	53.48	49.08	52.17	51.52	56.83	<0.001
PROMIS diarrhea T-score	45.70	46.37	48.43	45.17	53.98	<0.001
PROMIS physical functioning T-score	50.44	49.85	46.84	53.19	42.24	<0.001
Arthritis diagnosis	23 (30%)	29 (46%)	64 (45%)	20 (21%)	94 (58%)	<0.001
Asthma diagnosis	12 (16%)	7 (11%)	29 (20%)	7 (7%)	49 (30%)	<0.001
Chronic obstructive pulmonary disease (COPD) diagnosis	1 (1%)	5 (8%)	6 (4%)	1 (1%)	19 (12%)	0.001
Diabetes diagnosis	4 (5%)	8 (13%)	22 (15%)	5 (5%)	36 (22%)	0.004
Upper gastrointestinal disease diagnosis	19 (25%)	15 (24%)	34 (24%)	12 (13%)	71 (44%)	<0.001
Depression diagnosis	17 (22%)	9 (14%)	53 (37%)	27 (28%)	79 (49%)	<0.001
Anxiety or panic disorder diagnosis	13 (17%)	8 (13%)	35 (24%)	20 (21%)	62 (38%)	<0.001
Degenerative disc disease diagnosis	10 (13%)	7 (11%)	29 (20%)	10 (11%)	57 (35%)	<0.001
History of pelvic pain	15 (20%)	3 (5%)	7 (5%)	11 (12%)	37 (23%)	<0.001
Sexual activity with the last month	41 (54%)	27 (43%)	54 (38%)	58 (61%)	56 (34%)	<0.001
History of hypertension	20 (26%)	21 (33%)	61 (43%)	23 (24%)	80 (49%)	<0.001
History of hyperlipidemia	14 (18%)	24 (38%)	44 (31%)	22 (23%)	67 (41%)	0.009
History of sleep apnea	9 (12%)	10 (16%)	21 (15%)	9 (9%)	45 (28%)	0.005
History of a psychiatric diagnosis	32 (42%)	11 (17%)	55 (38%)	41 (43%)	93 (57%)	<0.001
Past surgical procedure for LUTS	4 (5%)	6 (10%)	23 (16%)	11 (12%)	36 (22%)	0.008
No abnormal vaginal findings on physical exam	57 (75%)	35 (56%)	101 (72%)	78 (82%)	124 (77%)	0.004
No abnormal uterus findings on physical exam	50 (68%)	45 (71%)	83 (60%)	69 (73%)	87 (54%)	0.009
No notation of tenderness on physical exam	44 (59%)	50 (79%)	108 (77%)	84 (88%)	131 (81%)	<0.001
Average number of voids in 24 hours	7.6	7.5	8.4	7.1	8.6	0.0023
Average voided volume in 24 hours (mL)	1827.0	1717.5	1786.6	1786.6	1813.5	0.6728
Average number of intakes in 24 hours	6.5	6.4	6.3	6.3	6.4	0.7081
Average volume of intakes in 24 hours	1902.8	1682.9	1813.3	1813.3	1810.1	0.3028
Max voided volume	531.9	543.2	473.4	473.4	519.0	0.2134

https://doi.org/10.1371/journal.pone.0268547.t003

They most frequently reported "urgency with fear of leakage" (2.85), but did not report any substantial post-voiding symptoms. Women in cluster W4 (n = 95) reported multiple symptoms associated with stress urinary incontinence, as well as urgency urinary incontinence, and some post-void urinary incontinence. They were younger (mean 51 years), had less medical comorbidities (FCI = 1.24), and had a higher level of physical functioning (PROMIS T-score = 53.2) than others in the cohort. Women in W5 (n = 165) reported higher frequencies and severities of LUTS for all symptoms. For 27 out of 30 urinary symptoms, they reported the highest levels across all five clusters. These women were heavier (87.6 Kg), had the lowest level of physical functioning (PROMIS T-score = 42.2), had more medical comorbidities (FCI = 3.64), and more pregnancies (3.07) than the rest of the cohort. They also reported higher psychosocial difficulties in depression, anxiety, and perceived stress, as well as sleep disturbance. Clusters W2 and W4 had higher percent of Hispanic or Latino women than three other clusters.

The presence of multiple significantly different variables across the clusters demonstrates that clusters W1-W5 meet the concise definition of clustering given by Liao as: "The goal of

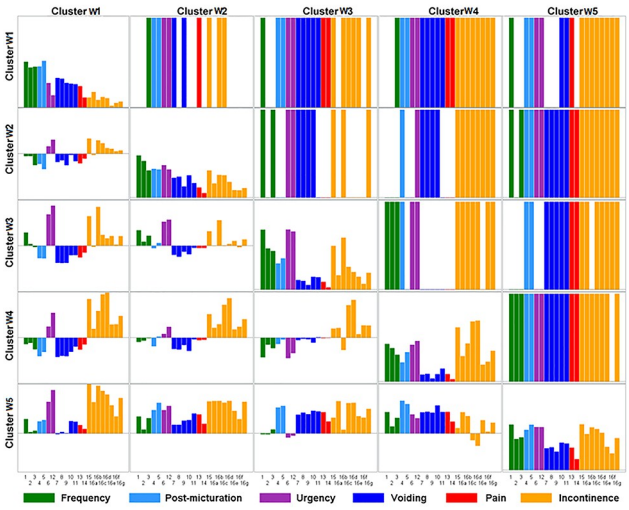
Cluster	N
W1	77
W2	64
W3	144
W4	95
W5	165

Fig 4. Radar plots illustrating mean values of urinary symptoms, demographics, clinical measurements, non-urinary PROs, comorbidities, and bladder diary variables for identified five clusters of women with LUTS. First row–urinary symptoms (LUTS Tool). Second row–clinical, non-urologic PRO, and demographic variables. Third row–comorbidities and anomalies identified by physical examination. Fourth row- bladder diary variables. Urinary symptoms are color-coded: Green = frequency; blue = post-micturition; purple = urgency; dark blue = voiding; red = pain; orange = incontinence.

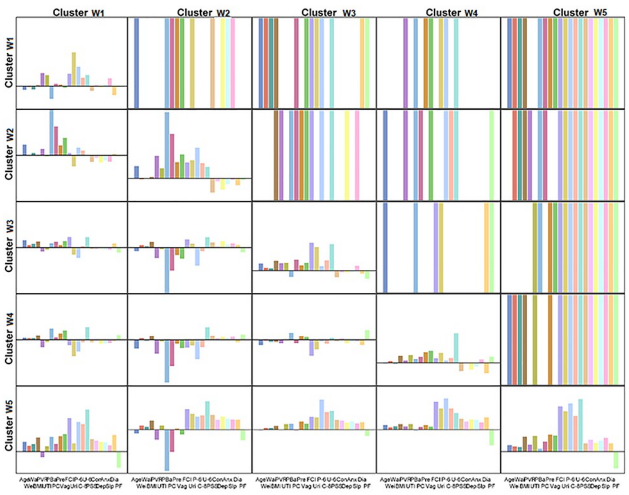
https://doi.org/10.1371/journal.pone.0268547.g004

clustering is to identify structure in an unlabeled data set by objectively organizing data into homogeneous groups where the within-group-object dissimilarity is minimized and the between-group-object dissimilarity is maximized" [8]. Fig 5 visually illustrates results of pairwise comparisons of the clusters in a matrix form. Fig 5A provides cluster comparisons by LUTS Tool variables. Elements on the diagonal of the matrix present the level of severity for each LUTS Tool question, i.e., the severity urinary symptom signature of the cluster. The triangle of boxes above the diagonal demonstrates variables significantly different in the pairwise comparison of the clusters; each colored bar indicates a significantly different variable. As seen, the majority of symptoms are significantly different in the pairwise comparison of the clusters. Elements in the lower triangle of the matrix present the difference in symptom severity levels; e.g., the first (upper) element in the triangle represents the difference between symptom severity levels in cluster W2 and cluster W1, indicating that urgency symptoms are more severe in cluster W2, while voiding and pain symptoms are more severe in cluster W1 than in cluster W2. Similarly, Fig 5B and 5C provide the results of pairwise comparison of the clusters for other variables from Tables 2 and 3, demonstrating multiple significantly different non-

LUTS Tool Comparison Across Clusters



Clinical Data Comparison Across Clusters



Comorbidity and Physical Exam Comparison Across Clusters

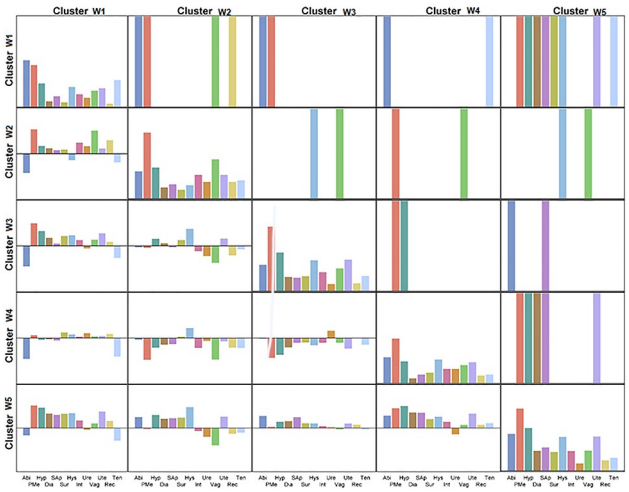


Fig 5. Results of the pairwise comparison of clusters W1-W5. (A) LUTS Tool variables. (B) Demographic and clinical variables. (C) Physical examination and comorbidities data. Boxes above the diagonal demonstrate significantly different variables in the pairwise comparison of the clusters. Each colored bar represents a significantly different variable. Boxes on diagonal are similar to radar plots and demonstrate the "signatures" of the clusters. Boxes below diagonal present the difference in the values of variables for each pair of clusters. Clusters are distinct and significantly different, not only by their urinary symptom signatures, but by multiple non-urologic variables, and comorbidities as well.

https://doi.org/10.1371/journal.pone.0268547.g005

urologic symptoms, physical examination, clinical and demographic variables. In summary, clusters are distinct and significantly different, not only by their urinary symptom signatures, but by multiple non-urologic variables as well. Importantly, these significant differences are demonstrated by omnibus test (Tables 2 and 3) and by pairwise comparison of the clusters (Fig 5).

Differential protein abundance in serum of women with LUTS versus non-LUTS controls

Fig 6 presents the volcano plots comparing abundances of 276 proteins in baseline serum samples of women with LUTS versus non-LUTS controls. Fig 6A compares the abundances for all 230 women with LUTS to 30 controls, while Fig 6B–6F provide similar comparisons for members of the identified clusters W1-W5 for whom proteomics data was available (n1 = 37, n2 = 38, n3 = 53, n4 = 42, n5 = 60). S3 Table in S1 File provides the lists of significantly

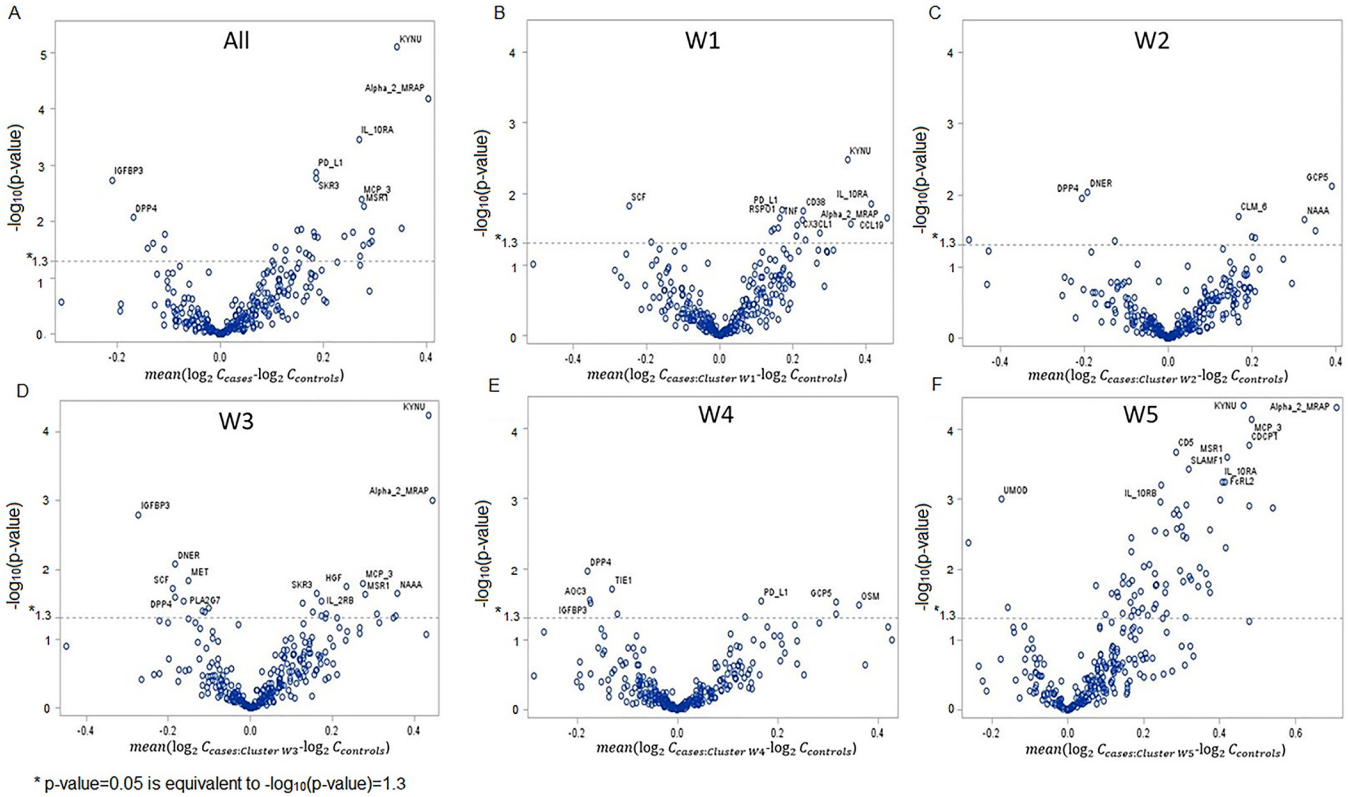


Fig 6. Volcano plots demonstrating differentially abundant proteins in women with LUTS vs. controls for 230 participants representing. (A) the whole cohort; and (B-F) for each of identified clusters W1-W5. Volcano plots allow for identification and visual representation of the differences in the data sets. Each small circle on volcano plots (A-F) represents mean abundance of one of 276 proteins in women with LUTS compared to non-LUTS controls. Horizontal axis represents mean fold-change on the logarithmic scale, while vertical axis represents p-value on the logarithmic scale. The higher the circle, the more significantly different its abundances in LUTS versus controls. The further the circle from zero on the horizontal axis, the larger the fold-change.

https://doi.org/10.1371/journal.pone.0268547.g006

Table 4. Proportion of significantly different variables in clusters W1-W5 and F1-F4.

	W1	W2	W3	W4	W5		F1	F2	F3	F4
W1		27%	46%	58%	64%	F1		34%	47%	58%
W2	27%		39%	37%	83%	F2	34%		27%	51%
W3	47%	27%		34%	66%	F3	47%	27%		39%
W4	58%	37%	34%		69%	F4	58%	51%	39%	
W5	64%	83%	66%	69%						
Mean					52.3%	Mean				42.7%

https://doi.org/10.1371/journal.pone.0268547.t004

differentially abundant proteins for each of the comparisons. Multiple differentially abundant proteins are observed in the serum samples of women with LUTS versus non-LUTS controls, both overall and between each cluster and controls. While some of these have been shown [72] to be associated with LUTS (e.g., tumor necrosis factor [TNF], interleukin-10 [IL-10], monocyte chemotactic protein [MCP], and transforming growth factor [TGF]), the remainders are novel. The highest number of the differentially abundant proteins of 70 (29 after FDR correction for multiple testing) is observed for cluster W5, which demonstrated the highest level of all urinary symptoms and comorbidities. Note that overlap between the lists of differentially abundant proteins is quite low, meaning that clusters W1-W5 are "biochemically" different. The highest overlap of 18 differentially abundant proteins is observed for cluster W5 and cluster W3, defined mainly by high urinary frequency, urinary urgency, and urge urinary incontinence. Interestingly, the lowest number of differentially abundant proteins (n = 10) are observed in clusters W2 (characterized by the presence of pelvic organ prolapse) and W4 (characterized by the presence of stress urinary incontinence), which are presumably driven by anatomic abnormalities, rather than biochemical changes. Without going into the detailed interpretation of these results, which are outside the scope of this paper, we think the observed differences in the differentially abundant proteins across W1-W5 clusters serve as important independent confirmation of the distinctiveness of the identified clusters.

Comparison of clusters W1-W5 with our previously published urinary symptom-based clusters F1-F4

Comparing quality of the clusters. Previously, we identified four clusters (F1-F4) by analyzing data on the same 545 women with LUTS using only urinary symptoms data collected via the LUTS Tool and AUA-SI (total of 52 variables) [41]. Since the same resampling procedure was performed when generating W1-W5 and F1-F4, both cluster structures are equally robust to the random variations of the cohort composition.

Distinctiveness, as determined by pairwise comparisons, was higher for the refined clusters compared with the previously published clusters (Tables 4 and 5). The proportion of significantly different variables in pairwise comparison of the clusters ranged from 27% to 83% (mean 52%) for the refined clusters, compared with a range of 27% to 58% (mean 43%) for the

Table 5. Proportion of core clusters in W1-W5 and F1-F4.

	PCC		PCC
W1	88%	F1	40%
W2	73%	F2	74%
W3	89%	F3	51%
W4	93%	F4	83%
W5	92%		
Mean	87%	Mean	62%

https://doi.org/10.1371/journal.pone.0268547.t005

previous clusters. The proportion of core clusters was also higher for W1-W5, compared with F1-F4; it ranged from 73% to 93% (mean 87%) for the refined clusters, compared with a range of 40% to 83% (mean 62%) for the previous clusters. Summarizing, there are more significantly different variables across refined clusters (W1-W5) than across our previously published clusters (F1-F4), and the refined clusters contain a higher percentage of core members for whom the probability to be in the given cluster is higher than the probability to be in all other clusters combined. Therefore, the refined clusters identified in the current paper by using additional urinary and non-urinary variables (total of 185 variables) are substantially more distinct than our previously published clusters based only on urinary symptoms measured by the LUTS Tool and AUA-SI (52 variables).

Comparing cluster membership. The Sankey diagram in Fig 7 serves to compare cluster membership in W1-W5 and in F1-F4. Refined cluster W1 is mostly composed of the members of cluster F1 without prolapse. Cluster W2 is formed by the members of cluster F1 with pelvic organ prolapse and includes some members of other clusters having prolapse and moderate urinary symptoms. Cluster W3 is mainly composed of members of F2 and F3 with urgency urinary incontinence. Cluster W4 is predominantly formed by members of F3 with stress urinary incontinence. Cluster W5 includes nearly all members of F4 and approximately 30% of F3 who have both urgency urinary incontinence and stress urinary incontinence symptoms.

Comparing radar plots. Fig 8 provides comparison of radar plots for the urinary symptom signatures of refined clusters W1-W5 and symptom-based clusters F1-F4. There are substantial similarities in the urinary symptom signatures of our previously published clusters, and of the refined clusters. Radar plots for W5 and F4 are similar in presenting all urinary symptoms at a uniformly high level. Signatures of W4 and F3 are similar in presenting the combination of stress urinary incontinence, urgency, and voiding dysfunction symptoms. W3 and F2 present urinary urgency, urgency urinary incontinence, and mild voiding problems. Symptom signatures of W1 and F1 are similar, presenting mostly voiding and post-micturition problems. The signature of cluster W2 presents mild LUTS and is mostly defined by clinically significant pelvic organ prolapse. The observed similarity of the clusters' LUTS signatures confirms that additional variables did not result in radical changes, but rather in incremental changes that allowed for identification of the refined clusters, which are built upon, but are more distinct and uniform than, our previously published ones. We believe this is further evidence of the stability of the identified clusters. Urinary symptom data captured by the LUTS Tool and AUA-SI provided the foundation for data-driven subtyping of LUTS, while the remaining urinary and non-urinary variables allowed for identifying refined clusters that differ not only by urinary symptoms, but by other PRO, demographic, and clinical variables as well (Table 3, Figs 5, 6B and 6C). Importantly, cluster refinement enhanced the distinctiveness and uniformity of the clusters, as well as the confidence in cluster membership, by increasing the overall proportion of core clusters from 62% to 87%.

Evolution of refined clusters W1-W5 in 3- and 12-month follow-up

Fig 9 presents radar plots of the urinary symptom signatures of refined clusters W1-W5 at 3- and 12-month follow-up. As seen, the shapes of the radar plots are conserved, while their areas representing overall severity of LUTS are decreased due to improvement in urinary symptoms for some participants. The percentage of improvers varied across the clusters (Table 6). Improvers were defined as having a ½ SD or greater improvement between baseline and 12 months on the calculated LUTS Tool Summary Score (including all 22 LUTS Tool severity variables). We view stability of the urinary symptom radar plot signatures' shapes as additional evidence of robustness of the identified clusters.

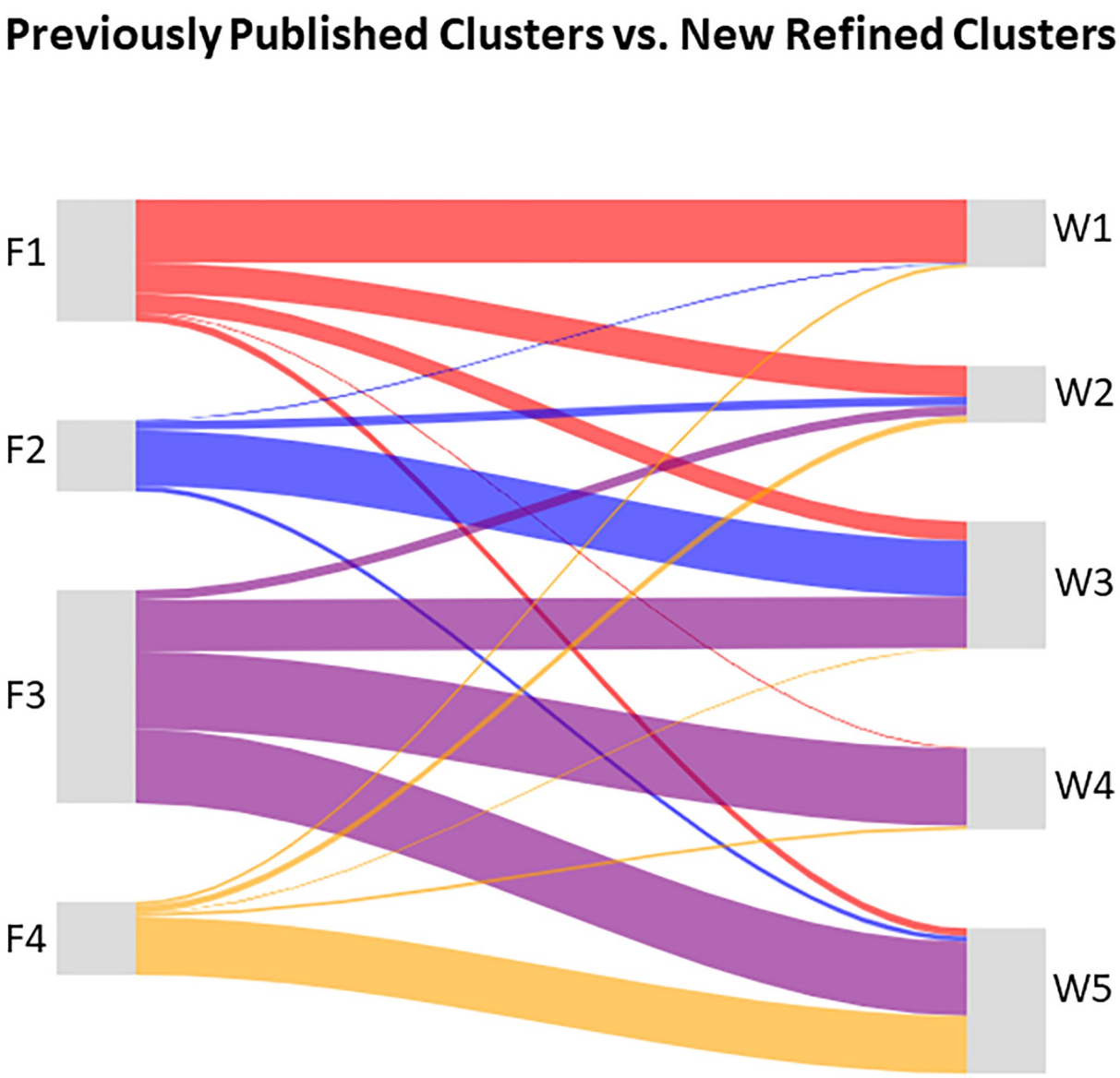
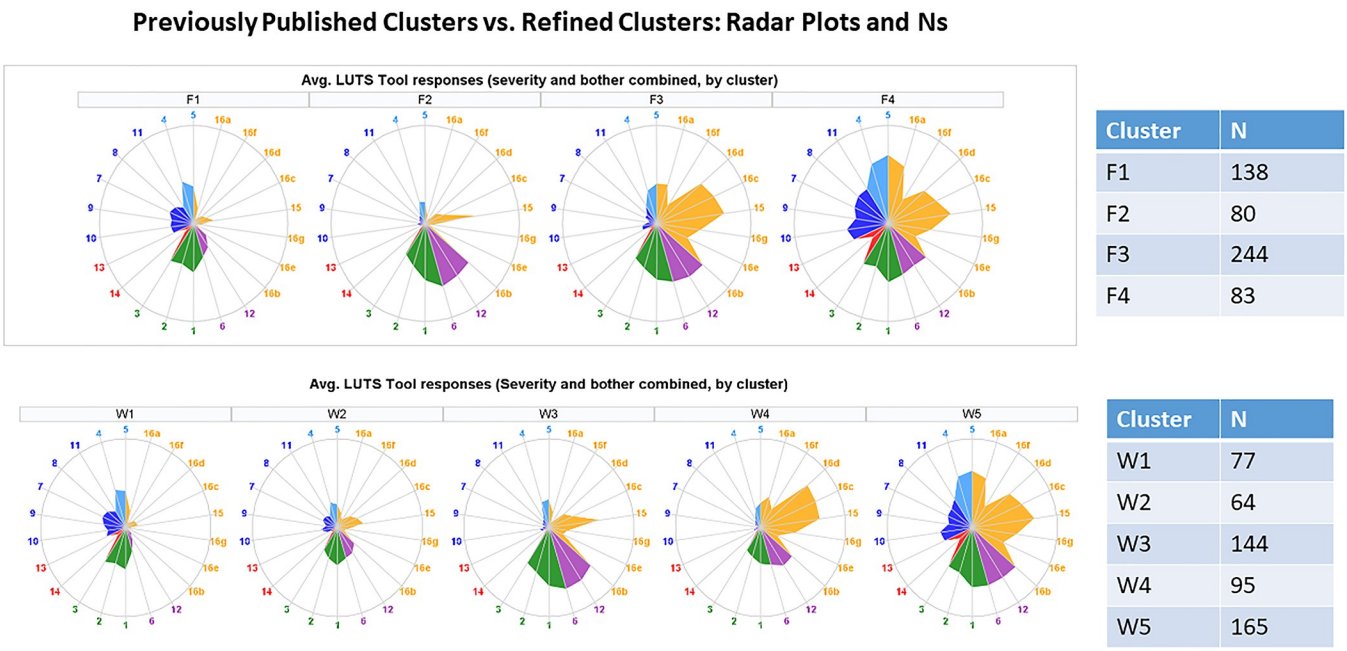


Fig 7. Sankey diagram comparing cluster memberships in W1-W5 and F1-F4. Cluster memberships in the refined cluster W1-W5 and previously published [41] urinary symptom-based clusters F1-F4 are compared. The new cluster W2 emerged, in which urinary symptoms are complicated by the presence of anterior vaginal wall prolapse. See text for more details on cluster comparison and properties of refined clusters W1-W5.

https://doi.org/10.1371/journal.pone.0268547.g007

Potential clinical significance of the identified clusters

The current paradigm for managing patients with LUTS is to assign a diagnosis based on a pre-defined symptom complex, such as overactive bladder (OAB), or based on a single predominant symptom, such as nocturia or stress urinary incontinence. Treatments are then administered based on these diagnoses [73,74]. Conventional classification of LUTS includes such partially overlapping groups as OAB wet, OAB dry, continent, stress urinary incontinence, urgency urinary incontinence, mixed urinary incontinence, underactive bladder, and bladder outlet obstruction. As stated in [39–41], there are limitations to this paradigm, as patients frequently present with multiple other urinary symptoms in addition to those being



Cluster	N
F1	138
F2	80
F3	244
F4	83

Cluster	N
W1	77
W2	64
W3	144
W4	95
W5	165

Fig 8. Comparison of radar plots of the urinary symptom signatures for clusters W1-W5 and F1-F4. Urinary symptom signatures (shapes of the radar plots) demonstrate pairwise similarities between the clusters F1-W1, F2-W3, F3-W4, F4-W5.

https://doi.org/10.1371/journal.pone.0268547.g008

treated, and these combinations of symptoms may be relevant to treatment selection. Diagnosis and treatment based solely on patients' chief complaints may be unsatisfactory, as they disregard other presenting symptoms. Mechanistic studies reveal that a functional impairment to a specific organ in the urinary tract may cause more than a single symptom. For example, a weak urethral sphincter is associated with both stress and urgency urinary incontinence [58,75]. This may explain why mixed incontinence is so common. This raises the question of how current diagnostic paradigms correspond with biological changes of the continence system and how symptoms occur in women seeking treatment, which was the main rationale for the unbiased data-driven subtyping of women with LUTS in LURN. As seen, none of the clusters W1-W5 identified in our analysis could be characterized by a single symptom, but rather by a combination of symptoms with various levels of severity, which are in concert with the clinical observations mentioned above [58,75]. The more detailed comparison of our refined clusters W1-W5 with the conventional LUTS groups and subtypes of LUTS identified by other researchers [36–38,76] is provided in Supplemental Material text in S1 File.

Clustering methodologies reported in studies on subtyping LUTS and other common diseases and disorders

Previous research on subtyping of diseases and disorders provides various levels of detail regarding the clustering methodologies that were employed. Below, we provide examples of high-quality studies using clustering methodologies for subtyping of common complex diseases, disorders, and symptom complexes within and outside of urologic domain. For instance, Coyne et al [36] provided detailed information on the 14 LUTS questions used for clustering and indicated that all the variables were scaled from 0 to 1. They took care of the robustness of clusters to the variation in composition of the cohort by performing clustering on the random 50% subset of participants first, and then by extending it to the whole cohort. They used k-means algorithm for clustering and scanned the number of clusters K from 3 to 7. They

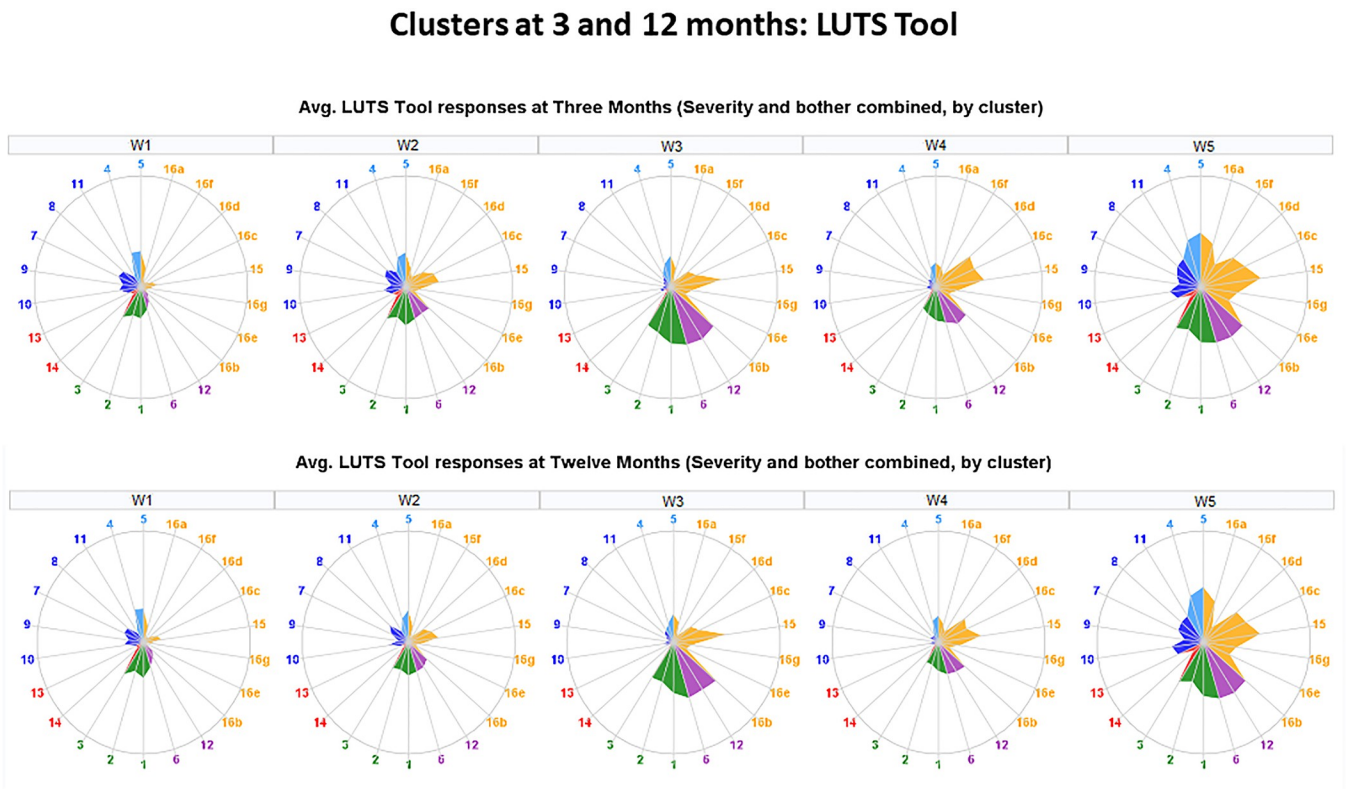


Fig 9. Evolution of the urinary symptom signatures in 3- and 12-month follow-up. First row–urinary symptom signatures for members of clusters W1-W5 at 3-month visit. Second row–urinary symptom signatures for members of clusters W1-W5 at 12-month visit. Note that the shape of the radar plots is conserved (similar to the radar plots in Figs 4 and 8), while the area of the radar plots is decreased due to symptoms improvement in some of the patients shown in Table 6.

https://doi.org/10.1371/journal.pone.0268547.g009

reported that the decision on the number of clusters was made by evaluating "each cluster model based on the clinical relevance and distinctiveness of each cluster, as well as the amount of variance accounted for by the cluster solution." However, they did not provide the names or values of criteria used for cluster evaluation. They provided detailed and informative descriptive statistics, but unfortunately, did not provide any tests on significantly different variables in pairwise comparison of the identified clusters.

Similarly, Hall et al [37] provided detailed information on the 14 LUTS questions and scaling. K-means clustering was performed with random 50% split of the cohort (split-half validation) similar to [36]. Detailed descriptive statistics and omnibus tests are provided across identified clusters and asymptomatic controls for significant difference in the variables not used for clustering, including demographics, comorbidities, risk factors, and lifestyle factors.

Table 6. Percentage of improvers in cluster W1-W5 in 12-month follow-up.

Cluster	N patients with 12-month follow-up	N (%) Improvers at 12 months
Cluster W1	50	23 (46%)
Cluster W2	51	23 (45%)
Cluster W3	106	57 (53%)
Cluster W4	70	49 (70%)
Cluster W5	119	69 (58%)

https://doi.org/10.1371/journal.pone.0268547.t006

However, no information on significant differences in the 14 LUTS variables used for clustering, and no information on pairwise comparison of the clusters is provided. Summarizing, these two LUTS clustering papers provide a reasonable level of details on scaling of the variables and on the clustering procedure, but unfortunately do not provide enough information on evaluation of the quality of the identified clusters.

In contrast, Miller et al [38] provided all necessary information on pairwise comparison of the identified clusters, both for six variables used for clustering, and for eight other variables collected but not used for clustering. Unfortunately, the authors did not perform any scaling of variables used for clustering. As stated in the paper, "the six clustering variables were (a) number of voids during daytime hours, (b) number of voids during nighttime hours, (c) daytime modal output volume in milliliters, (d) total 24-hour output volume in milliliters, (e) total 24-hour beverage intake in milliliters, and (f) BMI (a variable that was speculated to be related to intake)." If the volumes of related variables *c*, *d*, *e* were not scaled but entered into clustering in milliliters (values can be as high as 500), then these variables will provide the domineering contribution (compared with the number of daytime voids, typically < 20) to the 6-dimensional Euclidean distance between clustered objects, and will serve as drivers determining cluster membership. The contribution of these variables to the Euclidean distance would be much lower if they were entered in liters instead of milliliters, which would change the cluster membership. This is the problem indicated by Hair et al [62], with unscaled, unstandardized data of the inconsistency between cluster solutions when the scale of some variables is changed, which is a strong argument in favor of standardization. To our mind, the best solution of the problem is the use of the variables scaled by comparison with controls, as we described in the Methods section. Proper scaling is especially important when using heterogeneous data combining dimensionless and dimensional variables measured in different units, as in [38], where frequencies, volumes (mL), and BMI units (kg/m^2) are combined.

A broader look outside the LUTS domain shows that previous data on subtyping other common complex diseases provide different level of details on the clustering procedure and cluster evaluation as well. Table 7 below summarizes methodological information reported in the clustering papers [31–33] subtyping patients with asthma, diabetes, and sepsis, as well as in the LUTS papers [36–38] discussed above. More details on methodological information reported in [31–33] are provided in Supplemental Material text in S1 File.

Thoughts on minimal requirements for clustering publications

As shown above, publications using clustering for disease subtyping provide different levels of details on data preprocessing, clustering procedure, and cluster evaluation. In particular, information on scaling and weighting of the variables, values of criteria used for selection of the number of clusters, pairwise comparison of the clusters, and level of confidence in cluster membership are often not provided. This missing or hard-to-find information is important for better understanding of the papers' results, for comparison of the proposed phenotypes with previous and future classifications, and for potential refinement. With this in mind, and guided by the principles of FAIR data (findability, accessibility, interoperability, and reusability) [77,78], we think it is time for the clustering community to develop minimum requirements for clustering reports (MICRo), similar to minimum information about a proteomics experiment (MIAPE), developed by the proteomics community [79], and minimum information about a microarray experiment (MIAME), developed by the transcriptomics community [80]. We strongly believe that everybody, including the authors of this paper, would benefit from a collective consensus decision on the minimal information required for clustering publications.

Table 7. Methodological information provided in the clustering papers.

Paper	Variables	Preprocessing	Clustering algorithm	Number of clusters determination	Cluster evaluation
Moore et al [31]	34 variables derived from initial 726. Only list of 34 variables provided.	No info on scaling of 17 continuous variables; 17 composite variables ranked 0–10.	Agglomerative hierarchical clustering, Wards linkage.	Dendrogram demonstrates 5–6 groups. Five clusters selected due to small size of the sixth group. No other criteria.	Omnibus tests (analysis of variance [ANOVA], Kruskal-Wallis, chi-square) used on demographic, clinical, medication use, health care utilization, and biomarker variables. No pairwise comparisons of clusters.
Ahlqvist et al [32]	List of six variables used for clustering is provided.	Five variables standardized as z-scores. Presence of glutamic acid decarboxylase antibodies (GADA) binary variable.	Patients with GADA grouped into separate cluster. K-means with resampling for patients without GADA.	Schwarz's Bayesian criterion to determine number of clusters k = 4.	Box plots comparing 5 continuous variables used for clustering. Pairwise comparisons of clusters for multiple variables not used for clustering. Cluster validation in 3 independent cohorts.
Seymour et al [33]	List of 29 variables used for clustering is provided.	Variables standardized as z-scores.	Resampling-based consensus k-means clustering.	Consensus matrix heat map. Area under the CDF curve [34].	Pairwise comparison of variables clusters for variables used and not used for clustering. Validation in an independent cohort.
Coyne et al [36]	List and description of 14 EPIC LUTS questions used for clustering is provided.	Variables scaled from 0 to 1.	k-means clustering with split-half randomization. Values of k scanned from 3 to 8.	No names or values for criteria provided.	Descriptive statistics on variables both used and not used for clustering, but no tests for significance.
Hall et al [37]	List and description of 14 BACH LUTS questions used for clustering is provided.	Variables scaled from 0 to 1.	Hierarchical clustering and k-means clustering, split-half validation.	Pseudo F and t^2 statistics were used to determine number of clusters. However, values of statistics are not provided.	Omnibus tests (ANOVA, chi-square) on variables used and not used for clustering. No pairwise comparisons of clusters.
Miller et al [38]	List of six variables used for clustering is provided.	No explicit information on scaling.	Agglomerative hierarchical clustering, Wards linkage.	Dendrogram demonstrates 3–4 groups. Three groups selected based on visual examination of separation on canonical variables plane. No other criteria provided.	Pairwise comparison of clusters on six variables used for clustering and 8 variables not used for clustering. ANOVA for continuous and chi-square test for categorical data.

https://doi.org/10.1371/journal.pone.0268547.t007

Exact guidelines for the minimal requirements for clustering publications should result from the clustering community discussion. Here, we would like to call for such discussion and propose the below items that we believe are important for future guidelines:

1. Complete list of variables used for clustering.
2. Explicit information on scaling and weighting of clustering variables.
3. Clustering algorithm used (name of the function with options and parameter values, or code).
4. Exact definition of criteria used to determine the number of clusters. Values of criteria for the selected and alternative number of clusters. Preferably, more than one criterion should be presented.
5. Results of pairwise comparison of the clusters, with the indication of clinically meaningful and significantly different variables in the pairwise comparison–for all variables used for clustering, and for selected important variables not used for clustering (e.g., demographics).
6. Information on the level of confidence in cluster membership.

Limitations of the current study

Our paper carries some limitations. First, there are limitations in terms of the cohort, which included only treatment-seeking (e.g., potentially more difficult to treat patients) and predominantly white participants, some of whom (43%) received treatment prior to entering the study. Our analysis only contains women. Preliminary data analysis confirmed that sex is the major determinant of LUTS subtypes; therefore, sex-specific clustering was performed. We previously published the results of urinary symptom-based clustering of male participants [64]. Cluster refinement of male clusters, along the same lines as described in the current paper, is underway; the resulting refined subtypes will be compared with those found in the female cohort. Our control group used for scaling of the variables was relatively small (55 participants) but commensurate with the size of identified clusters. Not all data elements collected for the cases were available for the controls, so we used some literature data for general population and bladder diary data from a different study (EPI).

Second, there are limitations in terms of data elements used for clustering. Some objective measures that can be used in the diagnosis of LUTS in women, such as urodynamic testing, were not available. Urodynamics is clinically indicated in selected, but not all, LUTS patients due to the invasiveness of the procedure. Genomics data were not used so far. We do not expect that genomics data will produce dramatic effect on clustering results since LUTS is a highly prevalent common disease, especially in older age. Nevertheless, genotyping of the LURN participants is underway, which would allow for future cluster refinement by including the binary single nucleotide polymorphism (SNP) data using our novel weighted Tanimoto indices approach. Proteomics data are available for approximately 40% of the cohort and were not yet used for cluster refinement. However, they were used to demonstrate the presence of multiple significantly different proteins, indicating the refined symptom-based clusters are biochemically different.

Third, we developed methodology and a pipeline for integrating heterogeneous continuous and categorical data for clustering women with LUTS. We cannot claim that this is the preferable methodology for other data sets since data and research questions are different in different studies. However, we explicitly described our preprocessing and clustering procedures, as well as criteria used for determination of the number of clusters, cluster evaluation, and confidence level in cluster membership. We compared our methodology with alternative approaches and demonstrated that our methodology allows for combining heterogeneous continuous, categorical, and binary data, and that our refined clusters are more distinct than the previous urinary symptom-based clusters. Detailed description of the methodology, and comparison with the alternative approaches, allows interested readers to decide if it is suitable for their data and research questions. Availability of the pipeline source code allows for modifications, if needed.

Fourth, and most importantly, clinical significance of the identified clusters has yet to be determined. We already demonstrated the distinctness of our clusters, but now we need to establish their usefulness in clinical practice. This should be done through clinical trials, where treatments and outcomes of patients classified into the identified clusters would be compared with the standard treatment without the knowledge of cluster membership. Our preliminary analysis of 276 serum proteins in women with LUTS corroborated that the identified clusters are biochemically different. Further analysis of the affected biochemical pathways (potentially using even more comprehensive targeted proteomics assays, such as Olink Explore [3072 proteins] and/or new SomaScan assay [7000 proteins]), and their longitudinal dynamics as related to symptom trajectories, will follow with the goal to enhance understanding of different etiologies of the identified subtypes and potentially establishing more effective subtype-specific treatments. Further research will also include development of a software tool allowing for

classification of the "real-world" patient into one of the identified subtypes of LUTS and determination of the minimal set of variables sufficient for classification of patients with LUTS into identified subtypes in clinical practice. We hypothesize that the knowledge of cluster membership for a given patient would help clinicians to select an efficient treatment. This hypothesis could be tested in a study, where participants are randomized into two groups. The first group would be treated "as usual", while the second group would be classified into the identified subtypes of LUTS, with the information provided to the clinicians prescribing treatment. To clarify, we are not suggesting performing such a clinical study immediately. Cluster-specific treatments are yet to be determined. We believe that identification and further refinement of the LUTS subtypes with the omics and clinical data will improve our understanding of subtype etiologies and assist with identification of cause-specific and cluster-specific treatments. At that point, a clinical trial with cluster-specific treatments would be warranted, even if such treatments are identified not for all of the subtypes, but for some of them. We view subtyping not as a panacea, but as an important step in the development of personalized medicine.

Conclusion

A novel clustering pipeline for subtyping of common complex diseases, syndromes, and symptom complexes using heterogeneous continuous and categorical data was developed. The advantages of scaling variables by comparison with the controls without the disease of interest were discussed and illustrated by the simulated example. The novel weighted Tanimoto indices approach to integrate multiple binary variables into the clustering procedure was developed. A cluster refinement procedure using data available only for the subset of participants through semi-supervised clustering was proposed. A novel contrast criterion (CC) for resampling-based consensus clustering was proposed and compared with existing criteria for consensus clustering, i.e., consensus score (CS) and proportion of ambiguously clustered pairs (PAC). A simulated example demonstrated the advantages of CC over CS and PAC.

Information provided in the literature on subtyping common complex diseases and disorders was reviewed and shown to be often incomplete, especially with regard to data preprocessing, clustering procedures, and cluster evaluation. Suggestions for the minimum requirements for clustering publications were formulated, and the community effort to work on creating such requirements following the principles of FAIR data was called for.

Five distinct clusters of women with LUTS were identified by using 185 variables, including demographics, physical exam, LUTS and non-LUTS questionnaires, and bladder diary variables. The quality of the clusters was evaluated using established criteria (Calinski-Harabasz, Davies-Bouldin, Dunn, Point-Biserial, and Silhouette [22–26]), as well as novel contrast criterion (CC) and percentage of core members of the clusters (PCC). Distinctiveness of the clusters was confirmed by multiple significantly different variables in pairwise comparison of the clusters. Refined clusters W1-W5 were compared with our previously published urinary symptom-based clusters F1-F4, and were shown to be more distinct by having a higher percentage of significantly different variables and a higher percentage of the core members of the clusters. Importantly, targeted proteomics data confirmed that our refined clusters based on clinical data are biochemically different. Identification of the clinically and biochemically distinct subtypes of LUTS has provided a foundation for studies of subtype-specific etiologies and treatments. However, the results of the study should not be overgeneralized. Further refinement of subtypes is necessary and is coming both from new, more diverse cohorts (e.g., LURN 2 study) and from this cohort through inclusion of proteomic, genomic (grants ancillary to LURN), and neuroimaging data. Our paper provides methodology and a pipeline for such refinement and data integration.

Supporting information

S1 File. Supplemental material.
(DOCX)

Acknowledgments

Heather Van Doren, Senior Medical Editor with Arbor Research Collaborative for Health, provided editorial assistance on this manuscript.

The authors thank PLOS ONE's Academic Editor and Reviewers for their thorough and thoughtful feedback in the review process for this manuscript.

This is publication number 28 of the Symptoms of Lower Urinary Tract Dysfunction Research Network (LURN).

^The following individuals within the LURN Study Group were instrumental in the planning and conduct of this study at each of the participating institutions:

Duke University, Durham, NC. PIs: Cindy Amundsen, MD, Eric Jelovsek, MD; Co-Is: Kathryn Flynn, PhD, Todd Harshbarger, PhD, Jim Hokanson, PhD, Aaron Lentz, MD, David Page, PhD, Nazema Siddiqui, MD, Kevin Weinfurt, PhD Lisa Wruck, PhD; Study Coordinators: Yasmeen Bruton, Paige Green, Folayan Morehead

University of Iowa, Iowa City, IA. PIs: Catherine S Bradley, MD, Karl Kreder, MD, MBA, MSCE; Co-Is: Bradley A. Erickson, MD, MS, Daniel Fick, MD, Vince Magnotta, PhD, Philip Polgreen, MD, MPH; Study Coordinators: Mary Eno, Sarah Heady, Chelsea Poesch

Northwestern University, Chicago, IL. PIs: James W Griffith, PhD, Kimberly Kenton, MD, MS, Brian Helfand, MD, PhD; Co-Is: Carol Bretschneider, MD, David Cella, PhD, Sarah Collins, MD, Julia Geynisman-Tan, MD, Alex Glaser, MD, Christina Lewicky-Gaupp, MD, Margaret Mueller, MD; Study Coordinators: Sylwia Clarke, Melissa Marquez, Pooja Sharma, Michelle Taddeo, Pooja Talaty. Dr. Helfand and Ms. Talaty are at NorthShore University HealthSystem.

University of Michigan Health System, Ann Arbor, MI. PI: J Quentin Clemens, MD, FACS, MSCI; Co-Is: John DeLancey, MD, Dee Fenner, MD, Rick Harris, MD, Steve Harte, PhD, Anne P. Cameron, MD, Aruna Sarma, PhD, Giulia Lane, MD; Study Coordinators: Ashly Chimner, Linda Drnek, Emma Keer, Marissa Moore, Greg Mowatt, Sarah Richardson

University of Washington, Seattle, WA. PI: Claire Yang, MD; Co-I: Anna Kirby, MD; Study Coordinators: Lois Meryman, Brenda Vicars, RN

Washington University in St. Louis, St. Louis, MO. PI: H. Henry Lai, MD; Co-Is: Gerald L. Andriole, MD, Joshua Shimony, MD, PhD; Fuhai Li, PhD; Study Coordinators: Linda Black, Vivien Gardner, Patricia Hayden, Diana Wolff, Aleksandra Klim, RN, MHS, CCRC

Arbor Research Collaborative for Health, Data Coordinating Center, Ann Arbor, MI. PI: Robert Merion, MD, FACS; Co-Is: Victor Andreev, PhD, DSc, Brenda Gillespie, PhD, Abigail Smith, PhD; Project Manager: Melissa Fava, MPA, PMP; Clinical Monitor: Melissa Sexton, BA, CCRP; Research Analysts: Margaret Helmuth, MA, Jon Wiseman, MS, Jane Liu, MPH; Project Associate: Levi Hurley

National Institute of Diabetes and Digestive and Kidney Diseases, Division of Kidney, Urology, and Hematology, Bethesda, MD. Project Scientist: Ziya Kirkali MD; Project Officer: Christopher Mullins PhD; Project Advisor: Julie Barthold, MD

Author Contributions

Conceptualization: Victor P. Andreev.

Data curation: Margaret E. Helmuth, Abigail R. Smith.

Formal analysis: Victor P. Andreev, Margaret E. Helmuth, Gang Liu, Abigail R. Smith.

Funding acquisition: Victor P. Andreev, Robert M. Merion, Claire C. Yang, Cindy L. Amundsen, Brian T. Helfand, Catherine S. Bradley, John O. L. DeLancey, J. Quentin Clemens, H. Henry Lai.

Investigation: Victor P. Andreev, Margaret E. Helmuth, Gang Liu, Abigail R. Smith, Robert M. Merion, Claire C. Yang, Anne P. Cameron, J. Eric Jelovsek, Cindy L. Amundsen, Brian T. Helfand, Catherine S. Bradley, John O. L. DeLancey, James W. Griffith, Alexander P. Glaser, Brenda W. Gillespie, J. Quentin Clemens, H. Henry Lai.

Methodology: Victor P. Andreev, Margaret E. Helmuth, Gang Liu, Abigail R. Smith, J. Eric Jelovsek.

Project administration: Victor P. Andreev, Abigail R. Smith, Robert M. Merion, Claire C. Yang, Cindy L. Amundsen, Brian T. Helfand, Catherine S. Bradley, John O. L. DeLancey, J. Quentin Clemens, H. Henry Lai.

Resources: Victor P. Andreev, Robert M. Merion, Claire C. Yang, Cindy L. Amundsen, Brian T. Helfand, Catherine S. Bradley, John O. L. DeLancey, J. Quentin Clemens, H. Henry Lai.

Software: Victor P. Andreev, Margaret E. Helmuth, Gang Liu.

Supervision: Victor P. Andreev, Abigail R. Smith, Robert M. Merion, Claire C. Yang, Cindy L. Amundsen, Brian T. Helfand, Catherine S. Bradley, John O. L. DeLancey, J. Quentin Clemens, H. Henry Lai.

Validation: Victor P. Andreev, Margaret E. Helmuth, Gang Liu, Abigail R. Smith, Robert M. Merion, Claire C. Yang, Anne P. Cameron, J. Eric Jelovsek, Cindy L. Amundsen, Brian T. Helfand, Catherine S. Bradley, John O. L. DeLancey, James W. Griffith, Alexander P. Glaser, Brenda W. Gillespie, J. Quentin Clemens, H. Henry Lai.

Visualization: Victor P. Andreev, Margaret E. Helmuth, Gang Liu, Abigail R. Smith.

Writing – original draft: Victor P. Andreev, Margaret E. Helmuth.

Writing – review & editing: Victor P. Andreev, Margaret E. Helmuth, Gang Liu, Abigail R. Smith, Robert M. Merion, Claire C. Yang, Anne P. Cameron, J. Eric Jelovsek, Cindy L. Amundsen, Brian T. Helfand, Catherine S. Bradley, John O. L. DeLancey, James W. Griffith, Alexander P. Glaser, Brenda W. Gillespie, J. Quentin Clemens, H. Henry Lai.

References

1. Schadt EE, Lum PY. Reverse engineering gene networks to identify key drivers of complex disease phenotypes. J Lipid Res. 2006; 47:2601–2613. https://doi.org/10.1194/jlr.R600026-JLR200 PMID: 17012750
2. Becker KG. The common variants/multiple disease hypothesis of common complex genetic disorders. Medical Hypothesis. 2004; 62:309–317. https://doi.org/10.1016/S0306-9877(03)00332-3 PMID: 14962646
3. Relton CL, Davey Smith G. Epigenetic epidemiology of common complex disease: prospects for prediction, prevention, and treatment. PLoS Med. 2010; 7(10):e1000356. https://doi.org/10.1371/journal.pmed.1000356 PMID: 21048988
4. Coyne KS, Sexton CC, Thompson CL, Milsom I, Irwin D, Kopp ZS, et al. The prevalence of lower urinary tract symptoms (LUTS) in the USA, the UK and Sweden: results from the Epidemiology of LUTS (Epi-LUTS) study. BJU Int. 2009; 104(3):352–360. https://doi.org/10.1111/j.1464-410X.2009.08427.x PMID: 19281467
5. Irwin DE, Milsom I, Hunskaar S, Reilly K, Kopp Z, Herschorn S, et al. Population-based survey of urinary incontinence, overactive bladder, and other lower urinary tract symptoms in five countries: results of the

EPIC study. Eur Urol. 2006; 50(6):1306–1314; discussion 1314–1305. https://doi.org/10.1016/j.eururo.2006.09.019 PMID: 17049716

6. Litman HJ, McKinlay JB. The future magnitude of urological symptoms in the USA: projections using the Boston Area Community Health survey. BJU Int. 2007; 100(4):820–825. https://doi.org/10.1111/j.1464-410X.2007.07018.x PMID: 17550412
7. Coyne KS, Wein A, Nicholson S, Kvasz M, Chen CI, Milsom I. Economic burden of urgency urinary incontinence in the United States: a systematic review. J Manag Care Pharm. 2014; 20(2):130–140. https://doi.org/10.18553/jmcp.2014.20.2.130 PMID: 24456314
8. Liao TW. Clustering of time series data -a survey. Pattern Recognit. 2005; 38:1857–1874.
9. Duda RO, Hart PE, Stork DG. Pattern classification. 2nd Ed. New York: Wiley; 2001.
10. Robotti E, Manfredi M, Marengo E. Biomarkers discovery through multivariate statistical methods: A review of recently developed methods and applications in proteomics. J Proteomics Bioinform. 2014; S3:003.
11. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biology. 2002; 3(7):1–21. https://doi.org/10.1186/gb-2002-3-7-research0036 PMID: 12184810
12. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning, statistics. New York: Springer; 2001.
13. Wang L. Heterogeneous data and big data analytics. Autom Cont Inf Sci. 2017; 3(1):8–15.
14. Jain AK, Dubes RC. Algorithms for clustering data. Englewood Cliffs, NJ: Prentice Hall; 1988.
15. Giancarlo R, Scaturro D, Utro F. ValWorkBench: An open source Java library for cluster validation, with applications to microarray data analysis. Comp Meth Prog Biomed. 2015; 118:207–217. https://doi.org/10.1016/j.cmpb.2014.12.004 PMID: 25582071
16. Hartigan JA, Wong MA. Algorithm AS 136: A K-means clustering algorithm. J Royal Stat Soc Series C (Applied Statistics). 1970; 28(1):100–108.
17. Day WHE, Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. J Classif. 1984; 1:7–24.
18. Kohonen T. Self-organizing maps. Information sciences. Berlin: Springer; 1997.
19. Lum PY, Singh G, Lehman A, Ishkanov T, Vejdemo-Johansson M, Alagappan M, et al. Extracting insights from the shape of complex data using topology. Sci Rep. 2013; 3:1236. https://doi.org/10.1038/srep01236 PMID: 23393618
20. Bae E, Bailey J. COALA: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. International Conference on Data Mining 2006. Los Alamitos, CA, USA. IEEE Computer Society: 53–62.
21. Ramoni M, Sebastiani P, Cohen P. Multivariate clustering by dynamics. Proceedings of the 2000 National Conference on Artificial Intelligence (AAAI-2000). San Francisco, CA: 633–638.
22. Calinski T, Harabasz J. A dendrite method for cluster analysis. Comm Statistics. 1974; 3:1–27.
23. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. 1979; 1:224–227. PMID: 21868852
24. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J Royal Stat Society Series B. 2001; 63:411–423.
25. Rouseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Computational Applied Mathematics. 1987; 20:53–65.
26. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971; 66 (336):846–850.
27. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci. 1999; 96:6745–6750. https://doi.org/10.1073/pnas.96.12.6745 PMID: 10359783
28. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression. Science. 1999; 286(5439):531–537. https://doi.org/10.1126/science.286.5439.531 PMID: 10521349
29. Hayes DN, Monti S, Parmigiani G, Gilks CB, Naoki K, Bhattacharjee A, et al. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. J Clin Oncol. 2006; 24:5079–5090. https://doi.org/10.1200/JCO.2005.05.1748 PMID: 17075127
30. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci. USA 2004; 101:811–816. https://doi.org/10.1073/pnas.0304146101 PMID: 14711987

31. Moore WC, Meyers DA, Wenzel SE, Teague WG, Li H, Li X, et al. Identification of asthma phenotypes using cluster analysis in the severe asthma research program. Am J Respir Crit Care Med. 2010; 181:315–323. https://doi.org/10.1164/rccm.200906-0896OC PMID: 19892860

32. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018; 6:361–369. https://doi.org/10.1016/S2213-8587(18)30051-2 PMID: 29503172

33. Seymour CW, Kennedy JN, Wang S, Chang CCH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA. 2019; 321(20):2003–2017. https://doi.org/10.1001/jama.2019.5791 PMID: 31104070

34. Monti S, Tamayo P, Meserov J, Golub T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn 2003; 52:91–118.

35. Locke K Jr, Lai HH, Pontari MA, Clemens JQ, Kreder KJ, Krieger JN, et al. Discovery, validation, and novel visualization of subgroups in urologic chronic pelvic pain syndrome (UCPPS): Consensus clustering findings from the MAPP Research Network. J Urol. 2020; 203(4S):e104.

36. Coyne KS, Matza LS, Kopp ZS, Thompson C, Henry D, Irwin DE, et al. Examining lower urinary tract symptom constellations using cluster analysis. BJU Int. 2008; 101(10):1267–1273. https://doi.org/10.1111/j.1464-410X.2008.07598.x PMID: 18336611

37. Hall SA, Cinar A, Link CL, Kopp ZS, Roehrborn CG, Kaplan SA, et al. Do urological symptoms cluster among women? Results from the Boston Area Community Health Survey. BJU Int. 2008; 101 (10):1257–1266. https://doi.org/10.1111/j.1464-410X.2008.07557.x PMID: 18419699

38. Miller JM, Guo Y, Rodseth SB. Diary data subjected to cluster analysis of intake/output/void habits with resulting clusters compared by continence status, age, race. Nurs Res. 2011; 60(2):115–123.

39. Yang CC, Weinfurt KP, Merion RM, Kirkali Z, Lurn Study Group. Symptoms of Lower Urinary Tract Dysfunction Research Network. J Urol. 2016; 196(1):146–152. https://doi.org/10.1016/j.juro.2016.01.007 PMID: 26784646

40. Cameron AP, Lewicky-Gaupp C, Smith AR, Helfand BT, Gore JL, Clemens JQ, et al. Baseline lower urinary tract symptoms in patients enrolled in LURN: a prospective, observational cohort study. J Urol. 2018; 199(4):1023–1031. https://doi.org/10.1016/j.juro.2017.10.035 PMID: 29111381

41. Andreev VP, Gang L, Yang CC, Smith AR, Helmuth ME, Wiseman JB, et al. Symptom-based clustering of women in the Symptoms of Lower Urinary Tract Dysfunction Research Network (LURN) observational cohort study. J Urol. 2018; 200(6):1323–1331. https://doi.org/10.1016/j.juro.2018.06.068 PMID: 29990467

42. Coyne KS, Sexton CC, Kopp Z, Chapple CR, Kaplan SA, Aiyer LP, et al. Assessing patients' descriptions of lower urinary tract symptoms (LUTS) and perspectives on treatment outcomes: results of qualitative research. Int J Clin Pract. 2010; 64(9):1260–1278. https://doi.org/10.1111/j.1742-1241.2010.02450.x PMID: 20579138

43. Coyne KS, Barsdorf AI, Thompson C, Ireland A, Milsom I, Chapple C, et al. Moving towards a comprehensive assessment of lower urinary tract symptoms (LUTS). Neurourol Urodyn. 2012; 31(4):448–454. https://doi.org/10.1002/nau.21202 PMID: 22396308

44. Barry M, Fowler F Jr, O'Leary M, Bruskewitz RC, Holtgrewe HL, Mebust WK, et al. The American Urological Association symptom index for benign prostatic hyperplasia. The Measurement Committee of the American Urological Association. J Urol. 1992; 148(5):1549–1557. https://doi.org/10.1016/s0022-5347(17)36966-5 PMID: 1279218

45. Cheng J, Leng M, Li L, Zhou H, Chen X. Active semi-supervised community detection based on must-link and cannot-link constraints. PLoS ONE. 2014; 9(10): e110088. https://doi.org/10.1371/journal.pone.0110088 PMID: 25329660

46. Li Z, Liu J, Tang X. Pairwise constraint propagation by semidefinite programming for semi-supervised classification. Proceedings of the 25th International Conference on Machine Learning; Helsinki, Finland: 2008.

47. Şenbabaoğlu Y, Michailidis G, Li JZ. Critical limitations of consensus clustering in class discovery. Sci Rep. 2014; 4:6207. https://doi.org/10.1038/srep06207 PMID: 25158761

48. Ku JH, Jeong IG, Lim DJ, Byun S, Paick JS, Oh SJ. Voiding diary for the evaluation of urinary incontinence and lower urinary tract symptoms: Prospective assessment of patient compliance and burden. Neurourol Urodyn. 2004; 23(4):331–335. https://doi.org/10.1002/nau.20027 PMID: 15227650

49. Spiegel BM, Hays RD, Bolus R, Melmed GY, Chang L, Whitman C, et al. Development of the NIH Patient-Reported Outcomes Measurement Information System (PROMIS) gastrointestinal symptom scales. Am J Gastroenterol. 2014; 109(11):1804–1814. https://doi.org/10.1038/ajg.2014.237 PMID: 25199473

50. Pilkonis PA, Choi SW, Reise SP, Stover AM, Riley WT, Cella D, et al. Item banks for measuring emotional distress from the Patient-Reported Outcomes Measurement Information System (PROMIS®): depression, anxiety, and anger. Assessment. 2011; 18(3):263–283. https://doi.org/10.1177/1073191111411667 PMID: 21697139

51. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav. 1983; 24(4):385–396. PMID: 6668417

52. Yu L, Buysse DJ, Germain A, Moul DE, Stover A, Dodds NE, et al. Development of short forms from the PROMISTM sleep disturbance and sleep-related impairment item banks. Behav Sleep Med. 2011; 10 (1):6–24. https://doi.org/10.1080/15402002.2012.636266 PMID: 22250775

53. Clemens JQ, Calhoun EA, Litwin MS, McNaughton-Collins M, Kusek JW, Crowley EM, et al. Validation of a modified National Institutes of Health chronic prostatitis symptom index to assess genitourinary pain in both men and women. Urology. 2009; 74(5):983–987. https://doi.org/10.1016/j.urology.2009.06.078 PMID: 19800663

54. Barber MD, Chen Z, Lukacz E, Markland A, Wai C, Brubaker L, et al. Further validation of the short form versions of the Pelvic Floor Distress Inventory (PFDI) and Pelvic Floor Impact Questionnaire (PFIQ). Neurourol Urodyn. 2011; 30(4):541–546. https://doi.org/10.1002/nau.20934 PMID: 21344495

55. Bump RC, Mattiasson A, Bø K, Brubaker LP, DeLancey JO, Klarskov P, et al. The standardization of terminology of female pelvic organ prolapse and pelvic floor dysfunction. Am J Obstet Gynecol. 1996; 175(1):10–17. https://doi.org/10.1016/s0002-9378(96)70243-0 PMID: 8694033

56. Groll DL, To T, Bombardier C, Wright JG. The development of a comorbidity index with physical function as the outcome. J Clin Epidemiol. 2005; 58(6):595–602. https://doi.org/10.1016/j.jclinepi.2004.10.018 PMID: 15878473

57. Cameron AP, Wiseman JB, Smith AR, Merion RM, Gillespie BW, Bradley CS, et al. Are three-day voiding diaries feasible and reliable? Results from the Symptoms of Lower Urinary Tract Dysfunction Network (LURN) cohort. Neurourol Urodyn. 2019; 38(8):2185–2193. https://doi.org/10.1002/nau.24113 PMID: 31347211

58. DeLancey JO, Fenner DE, Guire K, Patel DA, Howard D, Miller JM. Differences in continence system between community-dwelling black and white women with and without urinary incontinence in the EPI study. Am J Obstet Gynecol. 2010; 202(6):584.e1–584.e12. https://doi.org/10.1016/j.ajog.2010.04.027 PMID: 20510959

59. Benjamini Y, Hochberg Y. Controlling the false discovery rate–a practical and powerful approach to multiple testing. J Royal Statistical Society, Ser B. 1995; 57(1):289–300.

60. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodol. 2001; 27(1):85–95.

61. Raghunathan TE, Solenberger PW, Berglund P, Van Hoewyk J. IVEware: Imputation and variance estimation software. Ann Arbor: University of Michigan, Institute for Social Research, Survey Research Center. 2000.

62. Hair JR, Anderson RE, Tatham RL, Black WC. Multivariate data analysis. Prentice- Hall Inc: Upper Saddle River, NJ; 1998.

63. Andreev VP, Gillespie BW, Helfand BT, Merion RM. Misclassification errors in unsupervised classification methods. Comparison based on the simulation of targeted proteomics data. J Proteomics Bioinform. 2016; S14:005. https://doi.org/10.4172/jpb.S14-005 PMID: 27524871

64. Liu G, Andreev VP, Helmuth ME, Yang CC, Lai HH, Smith AR, et al. Symptom-based clustering of men in the Symptoms of Lower Urinary Tract Dysfunction Research Network (LURN) observational cohort study. J Urol. 2019; 202(6):1230–1239. https://doi.org/10.1097/JU.0000000000000354 PMID: 31120372

65. Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery. 1998; 2(3):283–304.

66. Ng MK, Li MJ, Huang JZ, He Z. On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007; 29(3):503–507. https://doi.org/10.1109/TPAMI.2007.53 PMID: 17224620

67. Szepannek G, Aschenbruck R. k-prototypes clustering for mixed variable-type data. CRAN Repository 2021. Available at: https://cran.r-project.org/web/packages/clustMixType/clustMixType.pdf. Accessed 7/16/21.

68. SAS clustering action set: Clustering with the k-prototypes algorithm. SAS visual statistics programming guide. Available at: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/casactstat/casactstat_clustering_examples06.htm. Accessed 7/16/21.

69. Rogers DJ, Tanimoto TT. A computer program for classifying plants. Science. 1960; 132(3434):1115–1118. https://doi.org/10.1126/science.132.3434.1115 PMID: 17790723

70. Bilenko M, Basu S, Mooney RJ. Integrating constraints and metric learning in semi-supervised clustering. Proceedings of the 21st International Conference on Machine Learning (ICML). Banff, Canada. July 2004. Available at: https://www.cs.utexas.edu/~ml/papers/semi-icml-04.pdf. Accessed 8/6/21.

71. Huang H, Cheng Y, Zhao R. A semi-supervised clustering algorithm based on must-link set. In Tang C et al (Eds). ADMA 2008; LNAI 5139:492–499.

72. Siddiqui NY, Helfand BT, Andreev VP, Kowalski JT, Bradley MS, Lai HH, et al. Biomarkers implicated in lower urinary tract symptoms: systematic review and pathway analyses. J Urol. 2019; 202(5):880–889. https://doi.org/10.1097/JU.0000000000000257 PMID: 30925127

73. Syan R, Brucker BM. Guideline of guidelines: urinary incontinence. BJU International. 2016; 117(1):20–33. https://doi.org/10.1111/bju.13187 PMID: 26033093

74. AUA (American Urological Association) Guidelines. Available at: https://www.auanet.org/guidelines. Accessed 7/22/21.

75. DeLancey JO, Trowbridge ER, Miller JM, Morgan DM, Guire K, Fenner DE, et al. Stress urinary incontinence: relative importance of urethral support and urethral closure pressure. J Urol. 2008; 179 (6);2286–2290. https://doi.org/10.1016/j.juro.2008.01.098 PMID: 18423707

76. Rosen RC, Coyne KS, Henry D, Link CL, Cinar A, Aiyer LP, et al. Beyond the cluster: methodological and clinical implications in the Boston Area Community Health survey and EPIC studies. BJU Int. 2008; 101(10):1274–1278. https://doi.org/10.1111/j.1464-410X.2008.07653.x PMID: 18419700

77. Go FAIR. FAIR (Findable, Accessible, Interoperable, Reusable) principals for scientific data. Available at: https://www.go-fair.org/fair-principles/. Accessed 8/18/21.

78. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Scientific Data. 2016; 3:160018. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244

79. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK Jr., Jones AR, et al. The minimum information about a proteomics experiment (MIAPE). Nat Biotechnol. 2007; 25(8):887–893. https://doi.org/10.1038/nbt1329 PMID: 17687369

80. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet. 2001; 29(4):365–371. https://doi.org/10.1038/ng1201-365 PMID: 11726920