Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimorbidity Patterns, Disease Networks, and Phenotype Clusters in Vietnamese Children

Nhan Thi Ho
v.nhanht6@vinmec.com
Vinmec International Hospital
Michelle Hermiston
VinUniversity
Anh-Minh Nguyen
Vinmec International Hospital
Van-Quyet Nguyen
Vinmec International Hospital
Quoc-Anh Dao
Vinmec International Hospital
Thi-Linh-Chi Tran
Vinmec International Hospital
Nam-Phong Nguyen
Vinmec International Hospital
Ngoc-Quang Nguyen
Vinmec International Hospital
Tat-Cuong Do
Vinmec International Hospital
Nhat-An Pham
Vinmec International Hospital

Article

Keywords:

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10350120/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Multimorbidity in children remains poorly characterized in low- and middle-income countries (LMICs), with no prior large-scale network-analytic study in Vietnam. We aimed to describe the prevalence, structure, and demographic determinants of pediatric multimorbidity using a population-level school health database.

Methods

We analyzed 283,926 annual school health visits from 72,909 children aged 1 to 18 years attending a private school system in Vietnam (2020 to 2025). Multimorbidity was defined as ≥ 2 ICD-10 diagnoses per child. A six-layer multiplex network was constructed to identify disease hubs via betweenness centrality. Temporal trajectories were assessed with directed binomial tests, and multivariable logistic regression estimated adjusted odds ratios for comorbidity pairs. Association rules were extracted via Apriori and phenotype clusters were derived using Uniform Manifold Approximation and Projection (UMAP) followed by Leiden community detection.

Results

Multimorbidity was present in 53.7% of children, with higher burden in males (55.7% vs. 51.5%) and the 6 to 11-year age group (65.7%). The strongest co-occurrences involved blindness/low vision with protein-energy malnutrition (OR 6.71) and hyperalimentation (OR 8.73), constituting a visual-nutritional double burden. Hypertrophy of tonsils and dental caries emerged as universal disease hubs. Hyperalimentation preceding obesity was the highest-risk temporal sequence (RR 2.45). Seven phenotype clusters were identified (multimorbidity prevalence 39.3% to 69.5%), with disorders of refraction and dental caries present across all groups.

Conclusion

Multimorbidity is highly prevalent in this urban Vietnamese pediatric population. A nutritional-ophthalmological cluster and an upper airway-oral health cluster support integrated school-based screening and interventions in Vietnam and comparable LMICs.

What’s Known on This Subject

No large-scale, network-analytic study of pediatric multimorbidity has previously been conducted in Vietnam or comparable Southeast Asian LMIC settings. The patterns and magnitude of diagnostic co-occurrence across the pediatric spectrum have not been formally examined.

What This Study Adds:

In 72,909 urban children attending a private school system in Vietnam (2020 to 2025), 53.7% were multimorbid. Two clinically actionable clusters including a nutritional-ophthalmological pairing and an upper airway-oral health cluster support integrated screening and intervention within school health frameworks.

INTRODUCTION

In children, multimorbidity, defined as the co-existence of two or more distinct health conditions in a single individual, signals cumulative developmental risk, increases the complexity of clinical management, and challenges healthcare systems designed around single-disease episodes of care [1–3]. Multimorbidity in children is associated with increased school absenteeism, poorer educational attainment, and higher rates of unemployment in later life [4]. Despite this, the epidemiology of pediatric multimorbidity in low- and middle-income countries (LMICs) remains poorly characterized, with scoping reviews identifying major evidence gaps in sub-Saharan Africa and Southeast Asia and noting inconsistent consideration of hierarchical survey data in published studies [5,6]. This gap is particularly important in regions where the epidemiological transition, characterized by concurrent burdens of infectious, nutritional, and non-communicable diseases, creates conditions highly conducive to complex, multi-system multimorbidity patterns [7].

In Vietnam, rapid urbanization, significant socioeconomic heterogeneity, increasing screen exposure, and dietary transition have simultaneously sustained traditional burdens of undernutrition, perinatal conditions, and upper respiratory infections while accelerating the emergence of refractive disorders, obesity, and dental disease [8–10]. Existing Vietnamese pediatric literature has predominantly described single-disease prevalence. The nature and magnitude of diagnostic co-occurrence across the pediatric spectrum has not been formally examined.

This study analyses annual school health check electronic data from 72,909 unique children and adolescents (aged 1 to 18 years) attending a private school system in Vietnam from 2020 to 2025 with the following aims: (1) to characterize the overall multimorbidity burden; (2) to identify specific diagnosis pairs with the strongest co-occurrence associations at both patient level and visit level; (3) to map the co-occurrence network structure across demographic strata; (4) to characterize the temporal ordering of diagnosis pairs and identify forward-dominant disease trajectories; (5) to identify clinically actionable predictive relationships between diagnoses; and (6) to identify distinct multimorbidity phenotype clusters.

METHODS

Study design and setting

We conducted a retrospective analysis of annual school health check electronic deidentified data from 283,926 visits of 72,909 unique children and adolescents (aged 1 to 18 years) attending a private school system in Vietnam from 2020 to 2025. Children with at least one healthcare encounter during the study period were included. Routine child health examination (RCHE) visits without a diagnosis were retained for accurate denominator calculations but excluded from disease co-occurrence analyses to prevent artificial inflation of multimorbidity estimates. The study was approved by the Vinmec Ethical Committee (approval number 0231/2024/CN/HDDD VMEC). Informed consent was waived for retrospective analysis of deidentified data.

The study, all methods and report were carried out in accordance with the Declaration of Helsinki and STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) guideline. STROBE checklist with our study corresponding status was included in Supplementary Note.

Data sources and diagnosis classification

Diagnosis data were coded using ICD-10 and grouped into clinically meaningful categories. Patient demographic data (age, sex) and visit calendar year were included for stratified analyses.

Multimorbidity definition and prevalence

Multimorbidity was defined as two or more distinct diagnoses (excluding RCHE) within the same patient across the study period, categorized as 0, 1, 2, 3, or ≥ 4 distinct diagnoses per patient. Age-sex standardized prevalence was calculated using direct standardization with a reference population structure derived from the cohort, stratified by three age groups (1 to 5, 6 to 11, ≥ 12 years) and sex, computed per 10,000 population with variance estimates accounting for stratum-specific sample sizes.

Co-occurrence analysis

Patient-level and visit-level co-occurrence

Disease co-occurrence was analyzed at two levels: patient-level (any co-occurrence across all visits for a given patient) and visit-level (co-occurrence within individual health check visits). For each diagnosis pair, we calculated: (1) observed co-occurrence frequency (n_{AB}), (2) Jaccard index (J = nAB/(nA + nB-nAB)) [11], and (3) lift (ratio of observed to expected co-occurrence under independence) [12]. Statistical significance was assessed using Fisher's exact test, and adjusted for multiple comparisons using the Benjamini-Hochberg (BH) procedure to control the false discovery rate at $\alpha = 0.05$ [13].

Stratified co-occurrence analysis

Co-occurrence metrics were calculated separately for sex (male, female) and age groups (1–5, 6–11, ≥ 12 years) to identify demographic-specific patterns.

Network analysis

Single-layer disease networks

Disease co-occurrence networks were constructed separately for patient-level and visit-level data, with diagnoses as nodes and significant co-occurrence pairs (BH-adjusted $p < 0.05$, lift > 1, n ≥ 5) as weighted edges. Edge weights were set to Jaccard index values. Networks were visualized using the Fruchterman-Reingold force-directed layout algorithm [14], with node size proportional to diagnosis prevalence using the igraph (v1.5.0) [15] and ggraph (v2.1.0) [16] packages.

Multiplex network analysis

To examine comorbidity patterns across demographic strata while maintaining a consistent node set, we constructed a multiplex network with six layers: sex-stratified patient-level networks (female, male), age-stratified patient-level networks (1 to 5, 6 to 11, ≥ 12 years), and overall visit-level network [17]. Betweenness centrality was normalized by dividing by (n-1)(n-2)/2. Universal disease hubs were identified by calculating a cross-layer hub score as the product of mean betweenness centrality across layers and the number of layers in which each diagnosis appeared as a significant node. Layer similarity was quantified using Jaccard overlap of edge sets between all pairs of layers using the multinet package (v4.1) [18].

Temporal disease trajectories

Temporal ordering of diagnoses was analyzed using a directed trajectory network [19]. For each patient with multiple diagnoses, we identified the first occurrence of each diagnosis. A directed edge from diagnosis A→B was created if A was diagnosed before B in at least one patient. Directional asymmetry was assessed using the binomial test with null

hypothesis of equal forward and reverse probabilities (p = 0.5), with BH correction for multiple testing. Forward relative risk was calculated as RR=[(nforward)/(nforward+nreverse)]/0.5.

Association rule mining

Association rules were extracted using the Apriori algorithm implemented in the arules package (v1.7-6) [20], with minimum support 0.005 (0.5%), minimum confidence 0.2 (20%), and minimum lift 1.5, and ranked by lift.

Adjusted comorbidity associations

Multivariable logistic regression models were fit for each diagnosis pair, modelling diagnosis B as outcome and diagnosis A as predictor, adjusting for sex, age group, calendar year, and study sites. Adjusted ORs with 95% confidence intervals were presented with p-values adjusted using the BH procedure [13]. Only diagnosis pairs with at least 10 co-occurring children were included to ensure model stability.

Multimorbidity phenotype clustering

For children with at least one diagnosis, we constructed a binary patient×diagnosis matrix for the top 20 diagnoses. UMAP dimensionality reduction was applied using the uwot package (v0.1.16) [21]. Multimorbidity phenotype clusters were identified using the Leiden algorithm applied to a k-nearest neighbor graph (k = 10) from the UMAP embedding, using the leiden package (v0.4.3) [22] with resolution parameter = 0.8.

All analyses and visualization were performed in R version 4.5.1 (R Foundation for Statistical Computing, Vienna, Austria) [23]. Statistical significance was set at α = 0.05 for all tests, with appropriate adjustments for multiple comparisons.

RESULTS

Study population

The final analytic dataset comprised 283,926 visits of 72,909 unique children with at least one health check visit during 2020–2025. Table 1 summarizes child characteristics. Children were distributed approximately evenly by sex (47.7% female, 52.3% male), with the 6 to 11 year age group comprising the largest proportion (41.9%). Among the total cohort, 58,559 children (80.3%) had at least one diagnosis.

Table 1
Study population characteristics and multimorbidity distribution overall and by sex and age group, 2020–2025.

Characteristic	Overall N = 72,909	Female N = 34,753	Male N = 38,156	Age 1 to 5 years N = 26,409	Age 6 to 11 years N = 30,536	Age ≥ 12 years N = 15,964
Sex, n (%)						
Female	34,753 (47.7%)			12,253 (46.4%)	14,447 (47.3%)	8,053 (50.4%)
Male	38,156 (52.3%)			14,156 (53.6%)	16,089 (52.7%)	7,911 (49.6%)
Children with ≥ 1 diagnosis	58,559 (80.3%)	27,769 (79.9%)	30,790 (80.7%)	17,874 (67.7%)	26,873 (88.0%)	13,812 (86.5%)
Multimorbidity category, n (%)						
0 diagnoses	14,350 (19.7%)	6,984 (20.1%)	7,366 (19.3%)	8,535 (32.3%)	3,663 (12.0%)	2,152 (13.5%)
1 diagnosis only	19,404 (26.6%)	9,856 (28.4%)	9,548 (25.0%)	7,348 (27.8%)	6,804 (22.3%)	5,252 (32.9%)
2 diagnoses	17,286 (23.7%)	8,467 (24.4%)	8,819 (23.1%)	5,490 (20.8%)	7,666 (25.1%)	4,130 (25.9%)
3 diagnoses	11,501 (15.8%)	5,246 (15.1%)	6,255 (16.4%)	3,085 (11.7%)	5,944 (19.5%)	2,472 (15.5%)
≥4 diagnoses	10,368 (14.2%)	4,200 (12.1%)	6,168 (16.2%)	1,951 (7.4%)	6,459 (21.2%)	1,958 (12.3%)
Multimorbid (≥ 2 diagnoses), n (%)	**39,155 (53.7%)**	**17,913 (51.5%)**	**21,242 (55.7%)**	**10,526 (39.9%)**	**20,069 (65.7%)**	**8,560 (53.6%)**
Distinct diagnoses per patient, median (range)†	2 (1 to 20+)	2 (1 to 18)	2 (1 to 20+)	1 (1 to 17)	2 (1 to 20+)	2 (1 to 19)

Multimorbidity: ≥2 distinct diagnoses. †Kruskal-Wallis test for distribution of distinct diagnoses per patient: p < 0.001 for both sex and age group.

Multimorbidity burden

Among children with at least one diagnosis (n = 58,559), multimorbidity (≥ 2 diagnoses) was present in 39,155 (66.9%). Considering the entire cohort of 72,909 children, 53.7% were multimorbid. The distribution of multimorbidity categories is shown in Table 1 and Fig. 1A. Multimorbidity burden differed significantly by sex and age group (Kruskal-Wallis p < 0.001). Males showed higher rates of multimorbidity (55.7%) compared with females (51.5%), and were substantially more likely to carry four or more diagnoses (16.2% vs 12.1%). By age group, the 6 to 11 year cohort demonstrated the highest multimorbidity burden (65.7% with ≥ 2 diagnoses), while the 0–5 year group showed the lowest burden (39.9%).

Multimorbidity trends by year

Multimorbidity rates rose from 22.1% in 2020 to 49.9% in 2025, with consistent male excess across all years (Table 2, Fig. 1B).

Table 2
Multimorbidity burden by index year and sex, 2020 to 2025.

Year	N children	N with ≥ 1 Dx	N (%) multimorbid (≥ 2 Dx)	% Female	% Male	MM rate Female (%)	MM rate Male (%)
2020	9,847	4,661	2,179 (22.1)	46.9	53.1	20.4	23.7
2022	9,793	6,435	4,505 (46)	47.6	52.4	43.5	48.2
2023	32,072	21,425	13,843 (43.2)	47.5	52.5	38.9	46.9
2024	32,787	23,025	11,804 (36)	47.6	52.4	31.7	39.8
2025	35,759	35,754	17,845 (49.9)	47.6	52.4	47.9	51.7

N children=total unique children with at least one visit in that year. MM rate=proportion of children with ≥ 2 distinct diagnoses. Data for 2021 not available.

Patient-level co-occurrence: top pairs by lift

Table 3 presents the ten strongest patient-level co-occurrence pairs ranked by lift, all statistically significant after BH correction (Figs. 2A, 3A). The strongest patient-level co-occurrence was between blindness and low vision and unspecified protein-energy malnutrition (lift 5.17, Jaccard 0.050, OR 6.71, $p < 0.001$), affecting 151 children. The second-strongest pair was other hyperalimentation with blindness and low vision (lift 4.83, Jaccard 0.154, OR 8.73, $p < 0.001$), affecting 938 children. The most frequently co-occurring pair by absolute count was disorders of refraction and dental caries (N = 13,926 children), though with small lift (0.96), reflecting high marginal prevalence of both conditions rather than specific etiological linkage (**Supplementary Materials**).

Table 3
Top 10 diagnosis pairs by patient-level lift.

Rank	Diagnosis A	Diagnosis B	Prev A n (%)	Prev B n (%)	N both	Lift	Jaccard	OR (Fisher)	Adjusted p-value*
1	Blindness and low vision	Unspecified protein-energy malnutrition	2503 (4.31)	678 (1.17)	151	5.173	0.0498	6.71	p < 0.001
2	Other hyperalimentation	Blindness and low vision	4512 (7.76)	2503 (4.31)	938	4.829	0.1544	8.73	p < 0.001
3	Other hyperalimentation	Obesity	4512 (7.76)	3321 (5.71)	1026	3.981	0.1507	6.58	p < 0.001
4	Impacted cerumen	Unspecified protein-energy malnutrition	989 (1.7)	678 (1.17)	44	3.815	0.0271	4.15	p < 0.001
5	Chronic inflammation of lacrimal passages	Acute laryngitis and tracheitis	5891 (10.13)	1880 (3.23)	658	3.454	0.0925	5.25	p < 0.001
6	Blindness and low vision	Impacted cerumen	2503 (4.31)	989 (1.7)	115	2.701	0.0341	3.02	p < 0.001
7	Hypertrophy of tonsils	Acute tonsillitis	10718 (18.44)	685 (1.18)	309	2.447	0.0279	3.71	p < 0.001
8	Obesity	Blindness and low vision	3321 (5.71)	2503 (4.31)	344	2.406	0.0628	2.82	p < 0.001
9	Other hyperalimentation	Impacted cerumen	4512 (7.76)	989 (1.7)	181	2.358	0.034	2.73	p < 0.001
10	Other hyperalimentation	Acute pharyngitis	4512 (7.76)	1226 (2.11)	211	2.218	0.0382	2.54	p < 0.001

*N = 72,909 children. Prev=prevalence among all children. Lift=observed/expected co-occurrence ratio. Jaccard index quantifying set overlap as J = nAB/(nA + nB−nAB). OR=odds ratio from Fisher's exact test. *Benjamini-Hochberg (BH)-adjusted p-values. All pairs significant after Benjamini-Hochberg false discovery rate correction.*

Visit-level co-occurrence

Visit-level analysis, examining co-occurrence within individual health check visits, identified patterns complementary to and partly distinct from the patient-level analysis (Fig. 2B). Table 4 presents the ten strongest visit-level pairs by lift. The strongest visit-level association was chronic inflammation of lacrimal passages with acute laryngitis and tracheitis (lift 7.25, Jaccard 0.093, OR 11.8, p < 0.001, N = 658 co-occurring visits), suggesting simultaneous diagnosis during acute upper respiratory presentations. The co-occurrence of blindness and low vision with protein-energy malnutrition or hyperalimentation also presented in the top 10 pairs (**Supplementary Materials**).

Table 4
Top 10 diagnosis pairs by visit-level lift.

Diagnosis A	Diagnosis B	Prev A n (%)	Prev B n (%)	N pairs	Lift	Jaccard	OR (Fisher)	Adjusted p-value
Chronic inflammation of lacrimal passages	Acute laryngitis and tracheitis	5892 (4.82)	1883 (1.54)	658	7.254	0.0925	11.821	< 0.001
Dentofacial anomalies [including malocclusion]	Diseases of tongue	4581 (3.75)	1536 (1.26)	243	4.224	0.0414	5.044	< 0.001
Gingivitis and periodontal diseases	Diseases of tongue	6154 (5.03)	1536 (1.26)	283	3.662	0.0382	4.42	< 0.001
Blindness and low vision	Unspecified protein-energy malnutrition	3311 (2.71)	854 (0.7)	81	3.504	0.0198	3.836	< 0.001
Other hyperalimentation	Blindness and low vision	8338 (6.82)	3311 (2.71)	707	3.132	0.0646	3.963	< 0.001
Gingivitis and periodontal diseases	Dentofacial anomalies [including malocclusion]	6154 (5.03)	4581 (3.75)	651	2.825	0.0646	3.378	< 0.001
Impacted cerumen	Acute tonsillitis	1015 (0.83)	697 (0.57)	16	2.766	0.0094	2.837	< 0.001
Vasomotor and allergic rhinitis	Impacted cerumen	1917 (1.57)	1015 (0.83)	44	2.766	0.0152	2.889	< 0.001
Acute pharyngitis	Other low birth weight	1245 (1.02)	848 (0.69)	23	2.665	0.0111	2.743	< 0.001
Obesity	Vasomotor and allergic rhinitis	5163 (4.22)	1917 (1.57)	170	2.101	0.0246	2.249	< 0.001

Significant pairs only (BH-adjusted p < 0.05, ≥ 5 co-occurrences). Lift > 1 indicates diagnoses co-occur more than expected by chance. Jaccard index quantifies co-occurrence similarity (0 to 1 scale). N pairs=number of visits with both diagnoses.

Multiplex network analysis and universal disease hubs

Multiplex network analysis across six demographic strata identified diagnoses serving as universal comorbidity hubs (Fig. 4). Acute tonsillitis exhibited the highest cross-layer hub score (1.10), indicating consistent high centrality across all strata, followed by dental caries (0.95) (Table 5, Figs. 5A, 5B). Network topology varied substantially across demographic strata, with Jaccard similarity of edge sets between layers ranging from 0.24 (age 0 to 5 vs ≥ 12 years) to 0.705 (age 6 to 11 vs male (Fig. 5C, **Supplementary Materials**).

Table 5
Top 10 universal disease hubs by cross-layer multiplex hub score.

Diagnosis	Hub score	N layers present	Mean betweenness	Mean degree
Acute tonsillitis	1.10	6	0.1837	4.17
Dental caries	0.95	5	0.1901	6.17
Impacted cerumen	0.75	6	0.1253	6
Hypertrophy of tonsils	0.65	6	0.1091	7.5
Other low birth weight	0.61	6	0.1023	2.5
Chronic rhinitis, nasopharyngitis and pharyngitis	0.42	6	0.0702	4.67
Blindness and low vision	0.35	6	0.0575	7.17
Diseases of tongue	0.34	6	0.0575	3.5
Unspecified protein-energy malnutrition	0.34	6	0.0565	3.5
Vasomotor and allergic rhinitis	0.30	6	0.0507	5.17

Hub score=mean betweenness centrality × N layers present. N layers=number of strata (maximum 6) in which the diagnosis appears as a significant network node. Mean betweenness and mean degree are averaged across layers.

Temporal disease trajectories

Directed trajectory analysis identified 95 significant directed disease pairs among 30 diagnosis nodes (BH-adjusted $p < 0.05$). Network centrality analysis identified gingivitis and periodontal diseases as the dominant trajectory endpoint (PageRank = 0.234, in-degree = 12), with dental caries as the second most central sink node (PageRank = 0.100, in-degree = 11), suggesting these two conditions represent common endpoints of multiple preceding disease pathways in children. Other hyperalimentation and chronic inflammation of lacrimal passages were the leading source nodes (out-degree = 13 and 11, respectively), initiating the greatest number of directed trajectories. The most clinically meaningful risk-elevating trajectory was other hyperalimentation preceding obesity (N = 626, RR = 2.45, $p < 0.001$). This is the only pair with a substantially supra-unity relative risk. Several trajectories showed near-perfect directional asymmetry, most notably chronic inflammation of lacrimal passages preceding disorders of refraction and accommodation (N = 3,884, asymmetry = 0.9997), indicating a virtually exclusive temporal ordering with negligible reverse-direction cases (Fig. 6, **Supplementary Materials**).

Adjusted comorbidity associations

Multivariable logistic regression models identified the strongest comorbidity associations controlling for demographic confounders (Table 6, Fig. 7). The highest adjusted ORs were observed for acute tonsillitis and hypertrophy of tonsils (bidirectional OR 4.34, 95% CI 3.72–5.07 and OR 4.33, 95% CI 3.71–5.05, both $p < 0.001$). Other strong associations included blindness and low vision preceding disorders of refraction (OR 4.36, 95% CI 3.92–4.84) and dentofacial anomalies associated with diseases of tongue (bidirectional ORs ≈ 3.0) (**Supplementary Materials**).

Table 6
Top 10 adjusted comorbidity associations from patient-level logistic regression.

Predictor diagnosis	Outcome diagnosis	Adj. OR	95% CI	N both	BH-adjusted p*
Acute tonsillitis	Hypertrophy of tonsils	4.34	3.72, 5.07	309	p < 0.001
Hypertrophy of tonsils	Acute tonsillitis	4.33	3.71, 5.05	309	p < 0.001
Blindness and low vision	Disorders of refraction and accommodation	4.36	3.92, 4.84	1,997	p < 0.001
Disorders of refraction and accommodation	Blindness and low vision	3.88	3.47, 4.34	1,997	p < 0.001
Acute pharyngitis	Other low birth weight	3.18	2.15, 4.68	28	p < 0.001
Other low birth weight	Acute pharyngitis	3.18	2.16, 4.80	28	p < 0.001
Diseases of tongue	Dentofacial anomalies [incl. malocclusion]	3.05	2.64, 3.52	261	p < 0.001
Dentofacial anomalies [incl. malocclusion]	Diseases of tongue	3.03	2.63, 3.50	261	p < 0.001
Impacted cerumen	Hypertrophy of tonsils	2.46	2.12, 2.84	288	p < 0.001
Redundant prepuce, phimosis and paraphimosis	Chronic rhinitis, nasopharyngitis and pharyngitis	2.47	2.16, 2.84	317	p < 0.001

*Adj. OR=adjusted odds ratio. CI = 95% confidence interval. N both=number of children with both diagnoses. Models adjusted for sex, age group, calendar year, and hospital. *Benjamini-Hochberg (BH)-adjusted p-values. All associations BH-adjusted p < 0.001.*

Sex differences in co-occurrence

Sex-stratified lift analysis revealed marked differences in comorbidity patterns (Fig. 3B). Females exhibited substantially stronger associations for metabolic-nutritional comorbidities: lift difference (female-male) was largest for other hyperalimentation × obesity (+ 1.86). Conversely, males demonstrated stronger associations for visual-nutritional comorbidity profiles.

Association rules

Association rule mining identified 110 significant sets, with the strongest associations clustering around nutritional and ophthalmological conditions. The highest-lift rule was the co-occurrence of disorders of refraction and accommodation with other hyperalimentation predicting blindness and low vision (lift = 6.46, confidence = 0.28, N = 764), followed by blindness and low vision predicting other hyperalimentation (lift = 4.83, N = 938) and obesity predicting other hyperalimentation (lift = 3.98, N = 1,026), collectively pointing to a coherent nutritional-ophthalmological multimorbidity cluster. A strong dental-oral cluster was also evident, with hypertrophy of tonsils, gingivitis, and dentofacial anomalies all showing consistent co-occurrence with dental caries (lift range: 1.13–1.50). Disorders of refraction and accommodation appeared as the most frequent consequent node across rules, reflecting its central role as a comorbidity hub in this pediatric population (Fig. 8, **Supplementary Materials**).

Multimorbidity phenotype clusters

Unsupervised UMAP dimensionality reduction and Leiden community detection identified seven distinct multimorbidity phenotype clusters among 58,559 children with ≥ 1 diagnosis (Fig. 9, Table 7). Clusters differed substantially in diagnostic composition, demographic profiles, and multimorbidity burden. The three large clusters shared broadly similar profiles, multimorbidity prevalence of 67 to 70%, and disorders of refraction and accommodation as the most prevalent condition (67 to 72%), but differed in their defining features. Cluster C2 was defined by enrichment for blindness and low vision (1.7×), other hyperalimentation (1.6×), and unspecified protein-energy malnutrition (1.6×), identifying a visual-nutritional phenotype. Cluster C3 was defined by chronic inflammation of lacrimal passages (1.4×) and acute laryngitis and tracheitis (1.2×), capturing a lacrimal-respiratory pattern. Despite their differing defining diagnoses, all clusters shared disorders of refraction and accommodation and dental caries as ubiquitous background conditions.

Table 7
Multimorbidity phenotype cluster characteristics (N = 58,559 children with ≥ 1 diagnosis).

Cluster	N children	% MM (≥ 2 Dx)	% age 1 to 5 years	% age 6 to 11 years	% age ≥ 12 years	% Female	Top defining diagnosis
C1	13,961	67.6	28.2	46.2	25.6	47.8	Disorders of refraction (72.3%), Dental caries (37.7%), Hypertrophy of tonsils (18.8%)
C2	21,503	69.5	31.8	46.8	21.4	48.2	Disorders of refraction (67.8%), Dental caries (38.2%), Hypertrophy of tonsils (19.5%)
C3	21,323	67.8	29.3	45.8	24.9	46.5	Disorders of refraction (69%), Dental caries (36.9%), Hypertrophy of tonsils (18.1%)
C4	28	39.3	7.1	50	42.9	60.7	Disorders of refraction (74.1%), Dental caries (25%), Obesity (10.7%)
C5	280	43.2	16.4	75.4	8.2	41.4	Disorders of refraction (81.8%), Dental caries (26.1%), Hypertrophy of tonsils (10.4%)
C6	22	40.9	27.3	31.8	40.9	54.5	Disorders of refraction (81.8%), Dental caries (31.8%), Redundant prepuce, phimosis and paraphimosis (9.1%), Dentofacial anomalies [including malocclusion] (9.1%)
C7	51	47.1	15.7	35.3	49	52.9	Disorders of refraction (80.4%), Dental caries (33.3%), Redundant prepuce, phimosis and paraphimosis (7.8%)

% MM=percentage with multimorbidity (≥ 2 diagnoses). % Female=percentage female within cluster. Top defining diagnoses show the three highest-prevalence diagnoses and their within-cluster prevalence. Clusters identified via UMAP embedding (Jaccard distance, n_neighbors = 15) and Leiden community detection (resolution = 0.8).

FIGURES & LEGENDS

DISCUSSION

This study provides the first large-scale, multi-year, network-analytic characterization of multimorbidity using annual school health check data of Vietnamese urban children aged 1 to 18 years.

Multimorbidity burden in context

The overall multimorbidity prevalence of 53.7% is consistent with a Chinese school health check study in Yangzhou (2019 to 2024, n = 22,512 students) which found a comparable prevalence of 53.9% when a similarly broad diagnostic scope was used,[1] and substantially higher than estimates from general population studies in Scotland (approximately 15 to 30%),[24] Zimbabwe (5 to 14% of adolescents) [2] and Kathmandu Nepal (40.7% of school-going adolescents) [3] reflecting both the broad diagnostic scope encompassing refractive disorders, dental conditions, nutritional diagnoses, and respiratory infections, and the specific disease ecology of Southeast Asian urban pediatric populations undergoing rapid epidemiological transition.

The male predominance in multimorbidity is consistent with broader patterns reported in school and community settings. A UK Millennium Cohort Study analysis found co-occurrence of adverse health conditions to be higher in boys and in lower-income households,[25,26] and the Yangzhou school health data found boys had higher multimorbidity rates than girls.[1] The age distribution is particularly informative: the highest multimorbidity burden in the 6 to 11 year group is consistent with school entry and early school years representing a period of concentrated diagnostic recognition.

The overnutrition-visual impairment double burden

The strongest patient-level (also confirmed by visit-level) association regarding other hyperalimentation with blindness and low vision is epidemiologically novel. The co-occurrence of these conditions (OR 8.0) is substantially stronger than what marginal prevalence alone predicts. This pairing is consistent with a shared pathway mediated by excessive screen-based activity: screen exposure simultaneously drives myopic progression and is associated with sedentary, high-caloric dietary patterns leading to overfeeding.

Universal disease hubs and multiplex network structure

Acute tonsillitis emerging as the single most universal hub (hub score 1.10) reflects its high prevalence and multiple biological connections across the upper airway, oral cavity, and immune pathways. Dental caries achieving the second highest hub score (0.95) across five of six demographic layers reinforces the argument that oral health represents a universal entry point for comorbidity in this population, with direct relevance to integrated school health program design. Hypertrophy of tonsils, while ranking fourth overall (hub score 0.65), appeared as a significant node across all six demographic layers, reflecting its role as a clinical gateway to ENT and dental diagnoses. The marked heterogeneity in network topology across demographic layers demonstrates that comorbidity architecture is substantially age- and sex-dependent suggesting age, sex specific screening and intervention strategies.

Temporal disease trajectories

Temporal trajectory analysis reveals that chronic inflammation of lacrimal passages and other hyperalimentation are the dominant source diagnoses, each initiating the greatest number of directed pathways. Disorders of refraction and accommodation, rather than being an entry point, predominantly appears as a destination, which is most often preceded by lacrimal inflammation, gingivitis, dentofacial anomalies, and blindness and low vision, with near-perfect directional asymmetry in several of these pairs. This likely reflects the structured nature of school health checks, in which ophthalmic assessment typically completes a cascade triggered by other clinical findings rather than initiating

it. The one biologically interpretable causal sequence (other hyperalimentation preceding obesity (RR = 2.45)) is consistent with the expected latency between dietary excess and manifest metabolic consequence,[19] and represents the only forward-dominant trajectory where the relative risk substantially exceeds unity.

Implications for integrated pediatric care in Vietnam

Our findings map directly onto two integrated care pathways actionable within the existing Vietnamese school health check framework. For the nutrition-vision cluster, co-located ophthalmological and nutritional screening during the same visit would capture the majority of the highest-lift pairs simultaneously. Screen-time counselling addressing the shared exposure pathway linking visual impairment and overnutrition could be targeted. For the upper airway-oral health cluster, co-located ENT and dental assessments and interventions should be recommended.

Strengths and limitations

Strengths of this study include the largest pediatric comorbidity analysis from Vietnam and among the largest from any LMIC school-health setting, the multi-method analytical framework enabling triangulation across complementary statistical approaches, and the dual patient-level and visit-level analysis providing both cumulative multimorbidity burden and simultaneous clinical co-presentation data.

Limitations include: (1) the retrospective design, which precludes causal inference for all findings except the temporal trajectory analysis; (2) the absence of individual sociodemographic variables (parental income, household screen time, dietary data) that would allow adjustment for the social determinants most relevant to the nutrition-vision cluster; (3) restriction to urban children attending a private school system, which may not be representative of the broader Vietnamese pediatric population; (4) potential variation in diagnostic coding specificity and completeness across hospitals and years; (5) the absence of mental health diagnoses, which scoping reviews have highlighted as a major component of pediatric comorbidity burden;[6] and (6) the increasing multimorbidity prevalence trend from 2020 to 2025 reflects in part the expansion of diagnostic coding completeness rather than a true epidemiological change.

CONCLUSIONS

Multimorbidity is highly prevalent in urban Vietnamese children attending private schools, with more than half carrying two or more diagnoses and the highest burden among males and primary school-age children. Two clinically actionable clusters emerge: a nutrition-vision double-burden cluster anchored by the overnutrition-visual impairment association, and a high-prevalence upper airway–oral health cluster. The hyperalimentation-to-obesity trajectory stands out as the only forward-dominant pathway, with implications for early nutritional intervention. Seven distinct phenotypes indicate etiological heterogeneity, requiring phenotype-aware rather than single-disease care models, providing an empirical foundation for multimorbidity-aware pediatric care in Vietnam and comparable LMIC settings.

Abbreviations

LMICs
low- and middle-income countries
ICD-10
International Statistical Classification of Diseases and Related Health Problems
10th Revision
UMAP
Uniform Manifold Approximation and Projection

OR
Odds ratio
RR
Risk ratio
95%CI
95% confidence interval
UK
United Kingdom
RCHE
Routine child health examination without a disease diagnosis
BH
Benjamini-Hochberg.

Declarations

Contributors

Nhan Thi Ho did conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing original draft, and writing review & editing.

Michelle Hermiston, Quang Ngoc Nguyen, Cuong Tat Do did supervision, writing review & editing.

Minh Anh Nguyen, Quyet Van Nguyen, Anh Quoc Dao, Chi Thi Linh Tran, Phong Nam Nguyen did data curation, resources, validation, writing review & editing.

An Nhat Pham did data curation, resources, supervision, validation, and writing review & editing.

All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

Declaration of interests

All authors declare no competing interests.

Data sharing

De-identified patient-level data cannot be shared publicly owing to institutional data protection policies. Aggregated data supporting the findings are available in the supplementary materials. Code for all analyses is available upon reasonable request to the corresponding author. No individual participant data will be made available.

Use of Artificial Intelligence

All scientific content, analyses, tables, figures and interpretations are the original work of the authors. The authors used AI-assisted tools for language editing and grammar checking during manuscript preparation.

References

1. Wang, J. et al. Prevalence of Multimorbidity Among School-Aged Children in the Yangzhou District of China. *Healthcare* ; 13. (2025).
2. Nzvere, F. et al. Multimorbidity Burden Among Adolescents From the Health and Wellbeing Screening Programme, Y-Check, in Chitungwiza, Zimbabwe. *Journal Adolesc. Health* (2024).
3. Tandon, K., Adhikari, N. & Pradhan, P. Co-occurrence of non-communicable disease risk factors and its determinants among school-going adolescents of Kathmandu Metropolitan City. *PLoS One* ; 17. (2022).
4. Fleming, M. et al. Neurodevelopmental multimorbidity and educational outcomes of Scottish schoolchildren: A population-based record linkage cohort study. *PLoS Med* ; 17. (2020).
5. Obasohan, P., Walters, S., Jacques, R. & Khatab, K. Risk Factors Associated with Multimorbidity among Children Aged Under-Five Years in Sub-Saharan African Countries: A Scoping Review. *Int J. Environ. Res. Public. Health* ; 20. (2023).
6. Romano, I., Buchan, C., Baiocco-Romano, L. & Ferro, M. Physical-mental multimorbidity in children and youth: a scoping review. *BMJ Open* ; 11. (2021).
7. Sharma, V. et al. A systematic review and meta-analysis estimating the population prevalence of comorbidities in children and adolescents aged 5 to 18 years. *Obes. Rev.* **20**, 1341–1349 (2019).
8. Nguyen, N. et al. The Updated Dual Burden of Malnutrition Among Vietnamese School-Aged Children: A Nationwide Cross-Sectional Study. *Nutrients* ; 17. (2025).
9. Long, K. Q. et al. Clustering Lifestyle Risk Behaviors among Vietnamese Adolescents and Roles of School: A Bayesian Multilevel Analysis of Global School-Based Student Health Survey 2019. *Lancet Reg. Health West. Pac.* **15**, 100225 (2021).
10. Mai, T. M. T. et al. The double burden of malnutrition in Vietnamese school-aged children and adolescents: a rapid shift over a decade in Ho Chi Minh City. *Eur. J. Clin. Nutr.* **74**, 1448–1456 (2020).
11. Jaccard, P. The distribution of the flora in the alpine zone. *New Phytol.* **11**, 37–50 (1912).
12. Brin, S., Motwani, R., Ullman, J. D. & Tsur, S. Dynamic itemset counting and implication rules for market basket data. In: *SIGMOD Record* : 255–264. (1997).
13. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc. B*. **57**, 289–300 (1995).
14. Fruchterman, T. M. J. & Reingold, E. M. Graph drawing by force-directed placement. *Softw. Pract. Exp.* **21**, 1129–1164 (1991).
15. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex. Syst.* **1695**, 1–9 (2006).
16. Pedersen, T. L. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. (2022). https://CRAN.R-project.org/package=ggraph
17. Kivelä, M. et al. Multilayer networks. *J. Complex. Netw.* **2**, 203–271 (2014).
18. Dickison, M., Magnani, M. & Rossi, L. *Multilayer Social Networks* (Cambridge University Press, 2016). 10.1017/CBO9781139941907
19. Jensen, A. B., Moseley, P. L. & Oprea, T. I. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat. Commun.* **5**, 4022 (2014).
20. Hahsler, M., Grün, B. & Hornik, K. arules -- A computational environment for mining association rules and frequent item sets. *J. Stat. Softw.* **14**, 1–25 (2005).
21. Melville, J. The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction. (2023). https://CRAN.R-project.org/package=uwot

22. Traag, V. & leiden R Implementation of Leiden Clustering Algorithm. (2023). https://CRAN.R-project.org/package=leiden
23. R Core Team. *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, 2025).
24. Barnett, K. et al. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* **380**, 37–43 (2012).
25. Hesketh, K., Law, C., Bedford, H. & Hope, S. Co-Occurrence of Health Conditions during Childhood: Longitudinal Findings from the UK Millennium Cohort Study (MCS). *PLoS One* ; 11. (2016).
26. Hesketh, K. et al. Co-occurrence and clustering of health conditions at age 11: cross-sectional findings from the Millennium Cohort Study. *BMJ Open* ; 6. (2016).

Figures

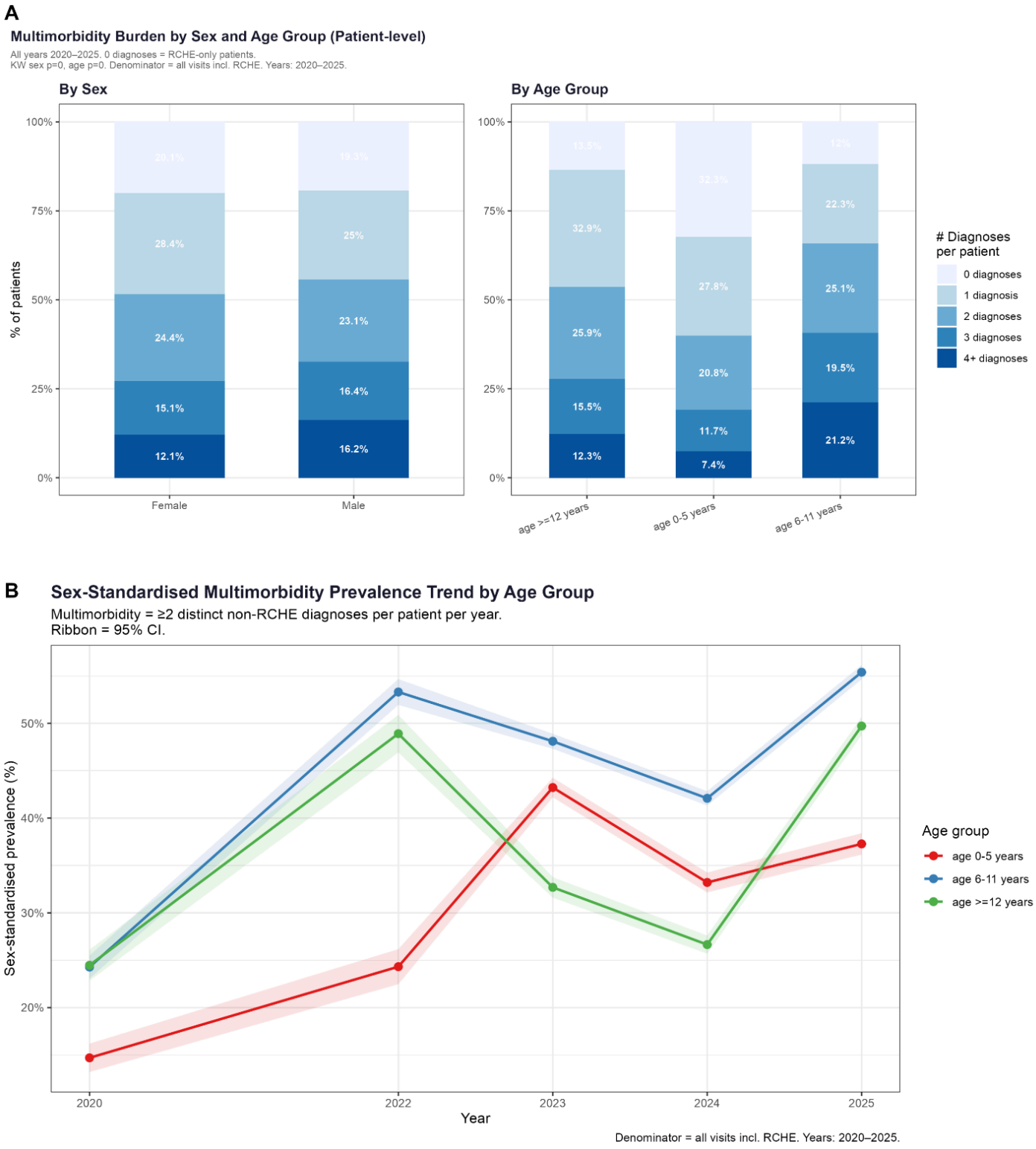


Figure 1

Multimorbidity burden by sex and age group, and sex-standardized prevalence trends by year.

Panel A. Multimorbidity Burden by Sex and Age Group (Patient-level).

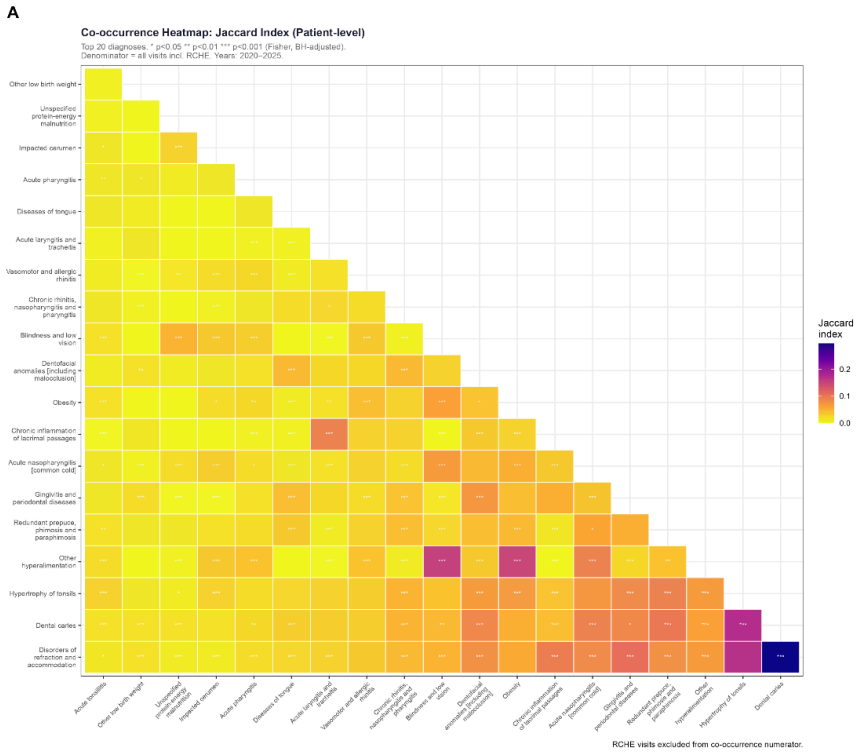
Stacked bar charts showing the percentage distribution of patients by number of distinct non-RCHE diagnoses (0, 1, 2, 3, ≥4), stratified by sex (left panel) and age group (right panel) across all years 2020–2025 (N=72,909 patients). Males showed higher multimorbidity burden than females (≥2 diagnoses: 55.7% vs 51.5%, ≥4 diagnoses: 16.2% vs 12.1%).

Among age groups, the 6–11 year group carried the highest burden (≥2 diagnoses: 65.7%, ≥4 diagnoses: 21.2%), followed by ≥12 years (53.6%, 12.3%) and 0 to 5 years (39.9%, 7.4%). Patients with 0 diagnoses were those attending RCHE visits only. Kruskal-Wallis tests: $p<0.001$ for both sex and age group.

Panel B. Sex-Standardized Multimorbidity Prevalence Trend by Age Group.

Line chart showing sex-standardized multimorbidity prevalence (≥2 distinct non-RCHE diagnoses per patient per index year) by age group across 2020 to 2025. Ribbon bands represent 95% confidence intervals. Multimorbidity prevalence rose from 22.1% in 2020 to 49.9% in 2025 overall, with the 6 to 11 year group consistently showing the highest prevalence at each time point. The non-linear trend across years reflects both true epidemiological change and increasing catchment size and diagnostic coding completeness over time (see Limitations). Denominator: all visits including RCHE. Years: 2020, 2022–2025 (2021 data not available).

Abbreviations: MM= multimorbidity (≥2 distinct diagnoses), RCHE= routine child health examination, CI= confidence interval.



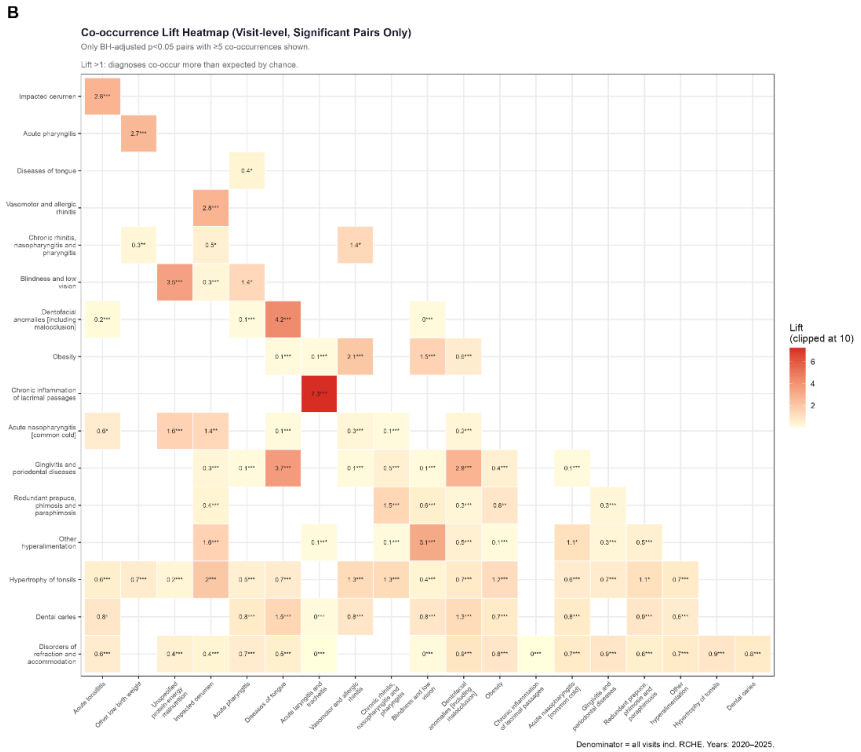


Figure 2

Patient-level and visit-level disease co-occurrence heatmaps for the top 20 diagnoses, 2020–2025.

Panel A. Co-occurrence Heatmap: Jaccard Index (Patient-level).

The lower-triangular heatmap displays pairwise Jaccard similarity indices for the top 20 diagnoses across 72,909 patients. Each cell represents one diagnosis pair, color intensity reflects the Jaccard index (range 0–1), where higher values indicate greater co-occurrence relative to union set size. Only pairs with Benjamini-Hochberg (BH)-adjusted

p<0.05 (Fisher's exact test) carry significance markers (* p<0.05, ** p<0.01, *** p<0.001). Diagnoses are ordered along both axes by overall prevalence. The denominator includes all patient visits including routine child health examinations (RCHE), RCHE visits are excluded from the co-occurrence numerator. Years: 2020 to 2025.

Panel B. Co-occurrence Lift Heatmap (Visit-level, Significant Pairs Only).

Lift values are shown only for BH-adjusted significant pairs (p<0.05) with ≥5 co-occurring visits. Lift >1 indicates diagnoses co-occur more frequently at the same visit than expected under statistical independence. Values are clipped at 10 for display, the strongest pair (chronic inflammation of lacrimal passages × acute laryngitis and tracheitis) had visit-level lift=7.25. Cells are blank where pairs did not meet significance or frequency thresholds. The denominator includes all visits including RCHE. Years: 2020–2025.

Abbreviations: BH= Benjamini-Hochberg false discovery rate correction, RCHE= routine child health examination.

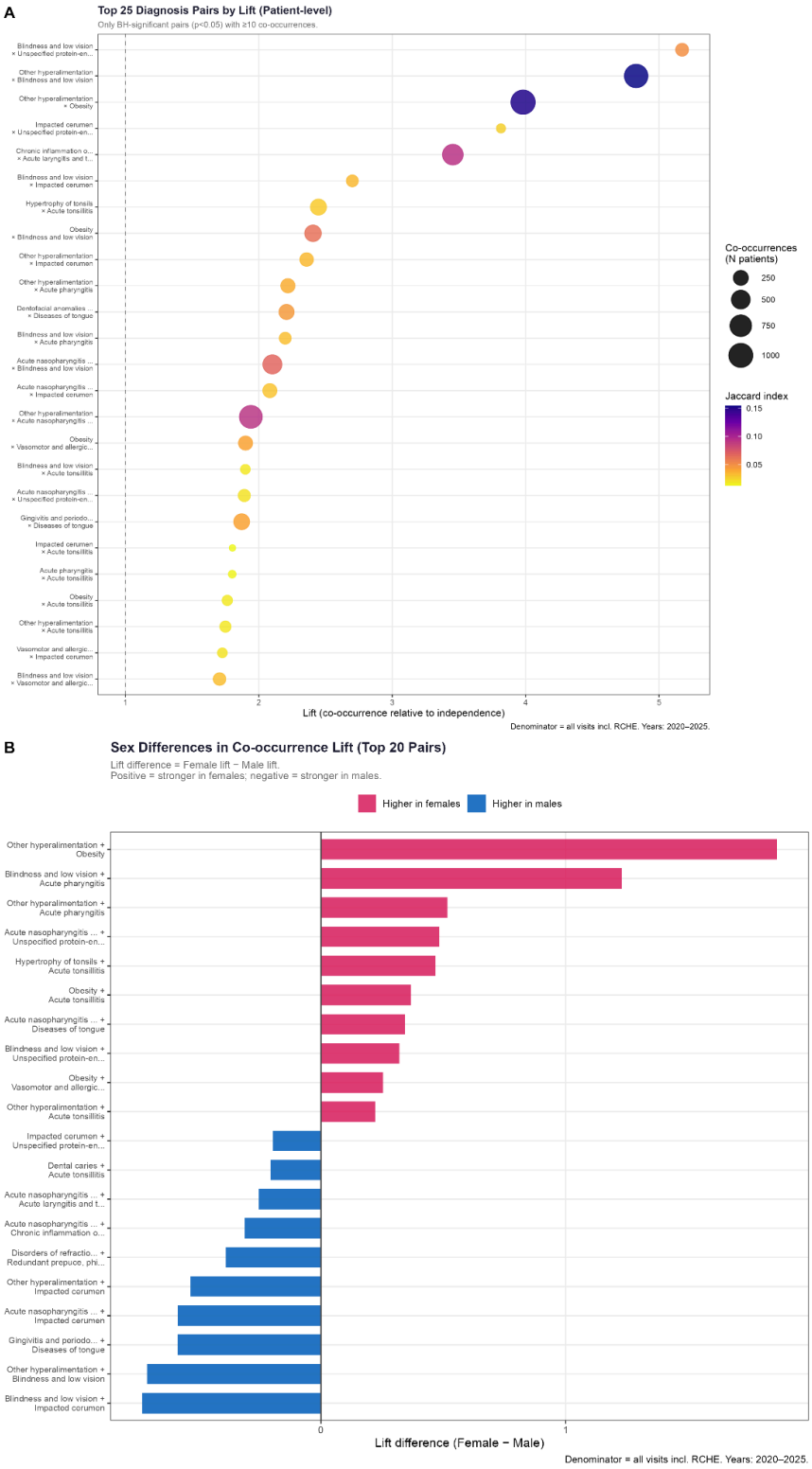


Figure 3

Top patient-level diagnosis pairs by lift and sex differences in co-occurrence strength.

Panel A. Top 25 Diagnosis Pairs by Lift (Patient-level).

Bubble chart showing the 25 BH-significant diagnosis pairs (p<0.05) with ≥10 co-occurrences, ranked by patient-level lift on the x-axis. Bubble area is proportional to the number of co-occurring patients (N patients), bubble color reflects the Jaccard similarity index. Lift values >1 indicate co-occurrence exceeds the frequency expected under

independence. All 25 pairs shown were statistically significant after BH correction. The dashed vertical line at lift=1 denotes the null expectation. The strongest pair was blindness and low vision × unspecified protein-energy malnutrition (lift=5.17, N=151), followed by other hyperalimentation × blindness and low vision (lift=4.83, N=938). Denominator: all visits including RCHE. Years: 2020 to 2025.

Panel B. Sex Differences in Co-occurrence Lift (Top 20 Pairs).

Horizontal bar chart showing the lift difference (female lift - male lift) for the 20 pairs with the largest absolute sex difference. Positive values (pink) indicate stronger co-occurrence in females, negative values (blue) indicate stronger co-occurrence in males. The largest female-predominant association was other hyperalimentation × obesity (lift difference +1.86), the largest male-predominant association was blindness and low vision × impacted cerumen (difference = 0.73). Lift values were computed separately within each sex stratum, the denominator for each stratum includes all visits for patients of that sex. Years: 2020 to 2025.

Abbreviations: BH= Benjamini-Hochberg, RCHE= routine child health examination.

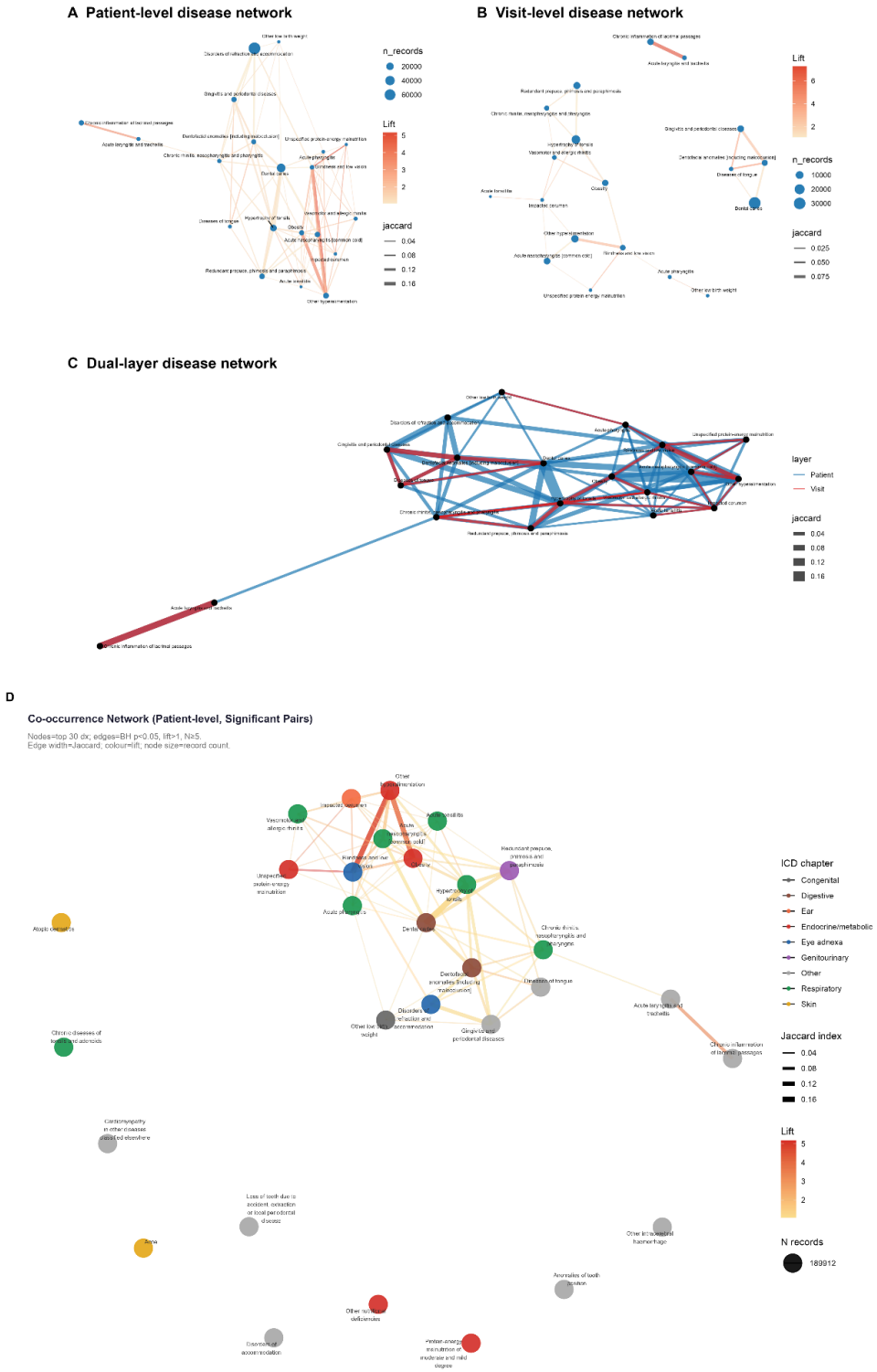


Figure 4

Disease co-occurrence networks at patient level, visit level, and combined dual-layer representation.

Panel A. Patient-level Disease Network.

Force-directed network (Fruchterman-Reingold algorithm) showing significant patient-level co-occurrence pairs (BH p<0.05, lift>1, N≥5). Nodes represent diagnoses, node size is proportional to the number of patient records. Edges

connect diagnosis pairs with significant co-occurrence, edge width is proportional to Jaccard index and edge color reflects lift (yellow=low, red=high). Only the top 20 diagnoses are displayed as nodes.

Panel B. Visit-level Disease Network.

As Panel A but restricted to co-occurrences within individual healthcare encounters (visit-level). Node size reflects visit-level record counts, edge widths use a separate Jaccard scale (0.025 to 0.075) reflecting the lower absolute co-occurrence frequencies at the visit level. The chronic inflammation of lacrimal passages–acute laryngitis and tracheitis dyad (lift=7.25) appears as the highest-lift edge in the network.

Panel C. Dual-layer Disease Network.

Overlay of patient-level (blue edges) and visit-level (red edges) networks, sharing the same node set. Edge width reflects Jaccard index (unified scale 0.04 to 0.16). The visual separation of the chronic inflammation of lacrimal passages–acute laryngitis dyad illustrates that this pair co-occurs with high lift at both levels but is spatially distinct from the main cluster, indicating epidemiological distinctness from the core multimorbidity network.

Panel D. Co-occurrence Network Patient-level, Significant Pairs (Top 30 Diagnoses).

Extended network including the top 30 diagnoses as nodes. Nodes are colored by ICD-10 chapter (see legend). Isolated peripheral nodes with no significant co-occurrence edges represent diagnoses present in the cohort but with no significant pairwise associations meeting the inclusion criteria (BH $p<0.05$, lift>1, $N\geq5$). Edge width reflects Jaccard index, edge color reflects lift.

Abbreviations: BH= Benjamini-Hochberg, ICD-10= International Classification of Diseases, 10th Revision.

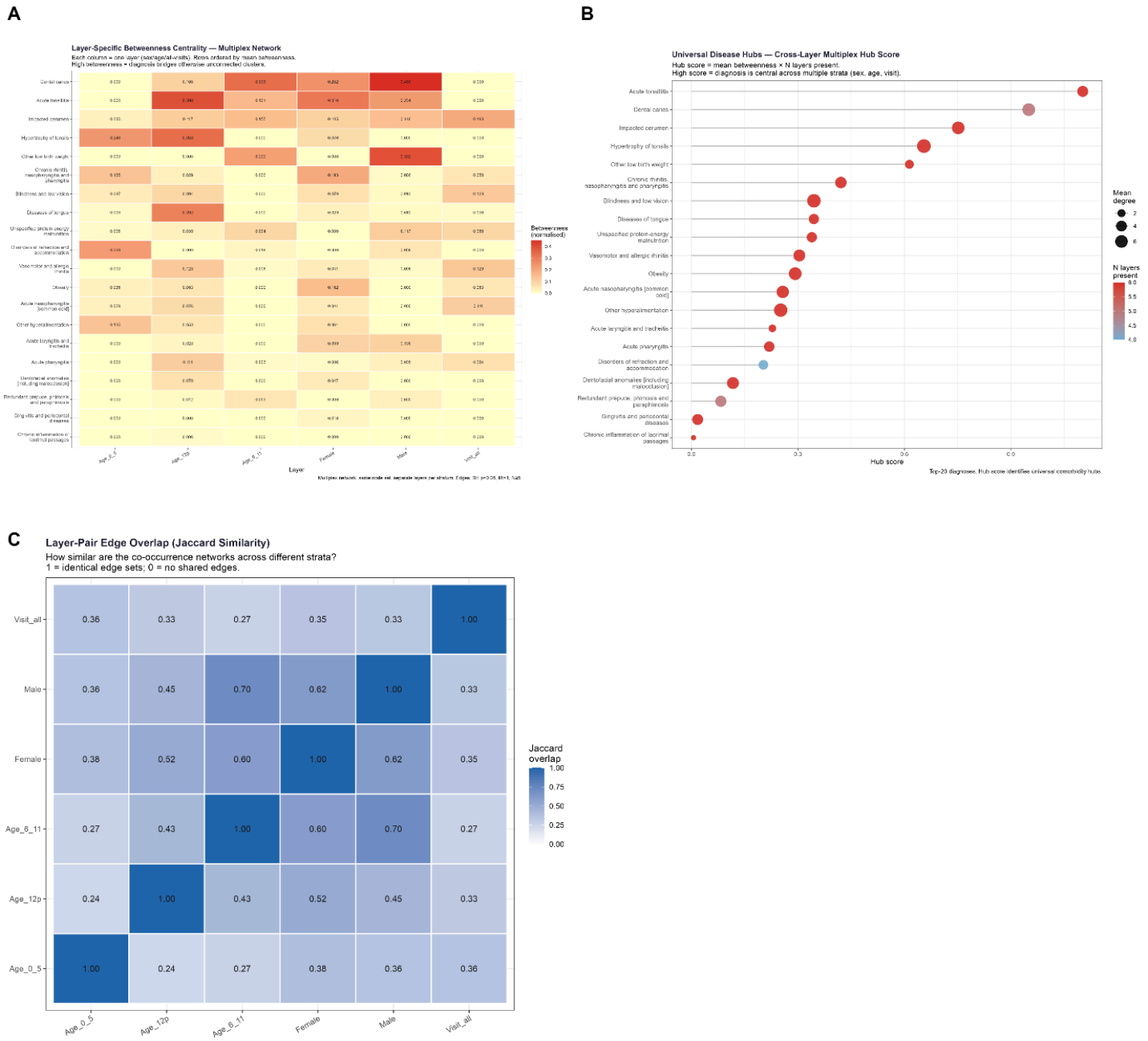


Figure 5

Multiplex network analysis: layer-specific betweenness centrality, universal disease hub scores, and cross-layer network similarity.

Panel A. Layer-Specific Betweenness Centrality: Multiplex Network.

Heatmap displaying normalized betweenness centrality for each diagnosis (rows) within each of six demographic network layers (columns): three age-stratified patient-level layers (Age_0–5, Age_12+, Age_6–11), two sex-stratified patient-level layers (Female, Male), and the overall visit-level layer (Visit_all). Rows are ordered by mean betweenness centrality across all layers. High betweenness centrality indicates a diagnosis frequently lies on the shortest path between other diagnosis pairs—functioning as a bridge connecting otherwise unconnected disease clusters. The

multiplex network used the same node set across all layers, edges within each layer were defined by BH $p<0.05$, lift>1, and N≥5.

Panel B. Universal Disease Hubs: Cross-Layer Multiplex Hub Score.

Lollipop chart ranking the top 20 diagnoses by cross-layer hub score (hub score=mean betweenness centrality × number of layers in which the diagnosis is present as a significant network node). Bubble size reflects mean degree across layers, bubble color reflects the number of layers present (scale 4 to 6). High hub scores identify diagnoses that are central across multiple demographic strata simultaneously. Hypertrophy of tonsils (hub score=1.10), acute tonsillitis (0.95), and obesity (0.75) were the top three universal hubs.

Panel C. Layer-Pair Edge Overlap (Jaccard Similarity).

Symmetric heatmap showing the Jaccard similarity of edge sets between all pairs of demographic network layers (J=1: identical co-occurrence structures, J=0: no shared edges). Jaccard overlap ranged from 0.24 (Age_0–5 vs Age_12+, indicating highly divergent network topology between the youngest and oldest age strata) to 0.71 (Age_6–11 vs Male, indicating the greatest similarity in comorbidity structure across any two strata).

Abbreviations: BH= Benjamini-Hochberg, Age_0–5= age 0 to 5 years, Age_6–11= age 6 to 11 years, Age_12+= age ≥12 years.

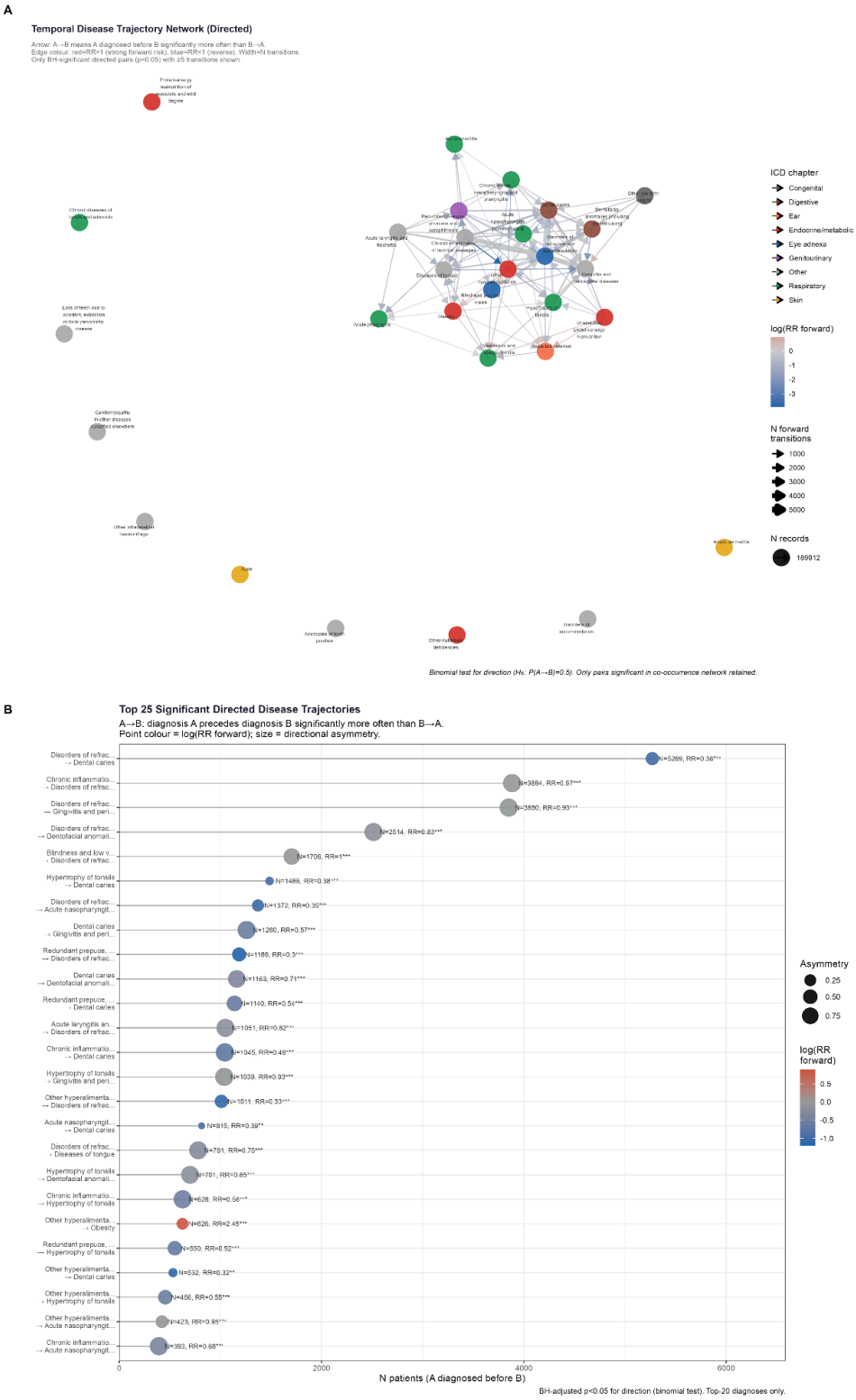


Figure 6

Temporal disease trajectory network and top 25 directed diagnosis sequences.

Panel A. Temporal Disease Trajectory Network (Directed).

Directed network in which each arrow A→B indicates that diagnosis A was recorded before diagnosis B significantly more often than the reverse (B→A) in the same patient, based on the first occurrence date of each diagnosis per patient. Edge color reflects log(RR_forward): red edges indicate RR>1 (strong forward risk) and blue edges indicate

RR<1 (reverse direction more likely). Edge width is proportional to the number of forward transitions. Node color reflects ICD-10 chapter, node size is proportional to total patient record count. Only BH-significant directed pairs (binomial test, p<0.05) with ≥5 transitions are shown. A total of 95 significant directed trajectories were identified. The only forward-dominant trajectory (RR>1) among the top 25 pairs by frequency was other hyperalimentation → obesity (RR=2.45, N=626), indicating that a hyperalimentation diagnosis substantially precedes and predicts a subsequent obesity diagnosis.

Panel B. Top 25 Significant Directed Disease Trajectories.

Lollipop chart ranking the 25 most frequent directed diagnosis pairs by number of patients in whom diagnosis A was recorded before diagnosis B (x-axis). Point color reflects log(RR_forward) and point size reflects temporal asymmetry (|n_forward - n_reverse| / [n_forward + n_reverse]). High asymmetry and large point size indicate strongly directional sequences. Disorders of refraction and accommodation preceded dental caries in 5,269 patients (RR=0.37), reflecting the high prevalence of refractive disorders as the entry diagnosis in this cohort. Analysis restricted to the top 20 diagnoses. BH-adjusted p<0.05 for directionality (binomial test).

Abbreviations: RR= relative risk, BH= Benjamini-Hochberg, ICD-10= International Classification of Diseases, 10th Revision.

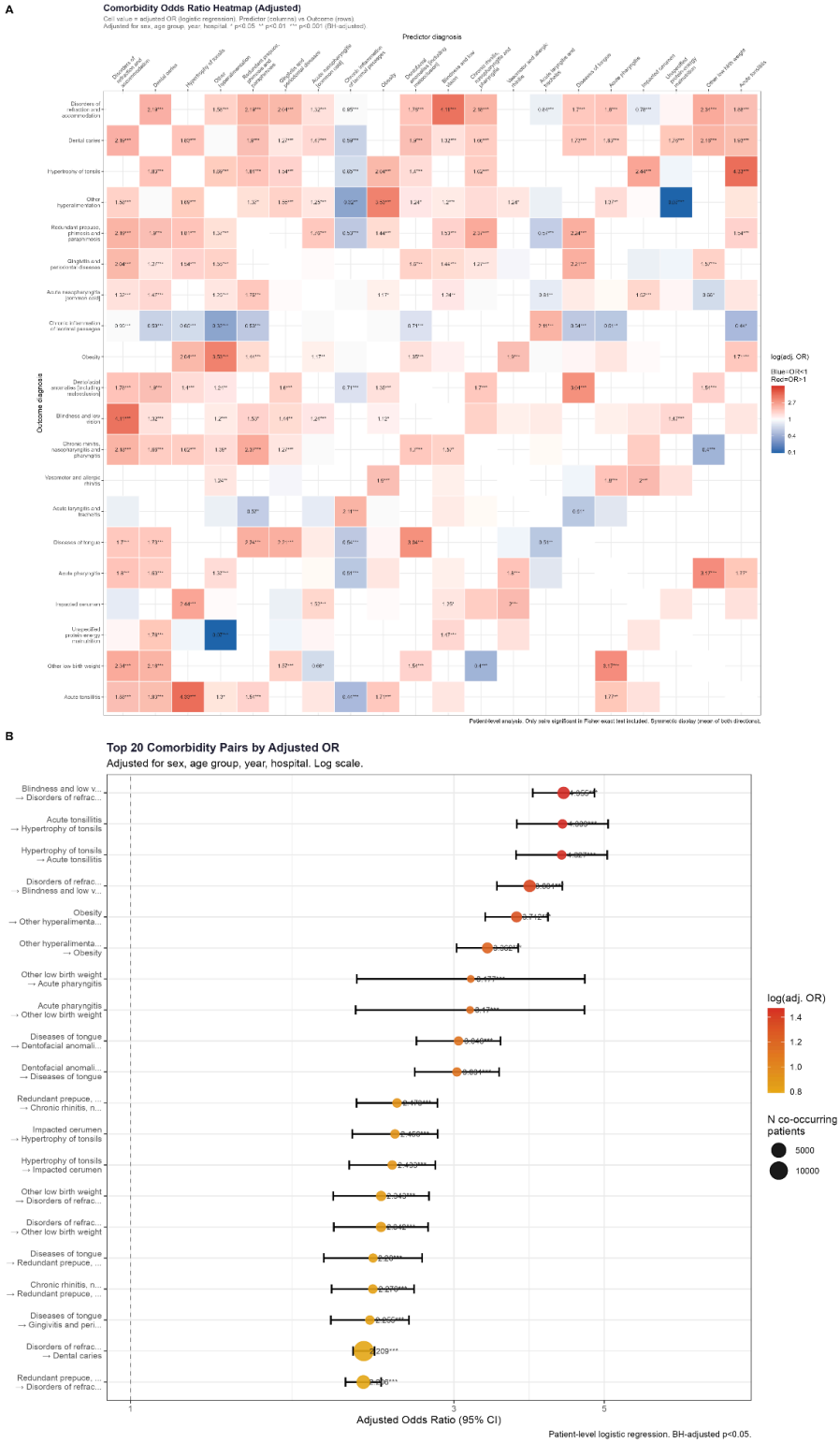


Figure 7

Adjusted comorbidity odds ratios from multivariable logistic regression.

Panel A. Comorbidity Odds Ratio Heatmap (Adjusted).

Heatmap of adjusted odds ratios (ORs) from patient-level logistic regression models predicting the presence of each outcome diagnosis (rows) from each predictor diagnosis (columns), adjusted for sex, age group, calendar year, and hospital. Cell values show adjusted OR, cells are colored on a log scale (blue: OR<1 [negative association], red: OR>1

[positive association], color intensity proportional to log[OR]). Only pairs significant in Fisher's exact test are displayed, the OR shown is the mean of both directional estimates for symmetric display. Significance markers: * p<0.05, ** p<0.01, *** p<0.001 (BH-adjusted). Blank cells did not meet the significance threshold or minimum co-occurrence criteria (N≥10).

Panel B. Top 20 Comorbidity Pairs by Adjusted OR (Forest Plot).

Forest plot on a log scale showing the 20 comorbidity pairs with the highest adjusted ORs. Each point represents one directional association (predictor → outcome), error bars show 95% confidence intervals. Point color reflects log(adjusted OR), point size is proportional to the number of co-occurring patients. The strongest associations were blindness and low vision → disorders of refraction (OR=4.36, 95% CI 3.92–4.84, N=1,997), acute tonsillitis → hypertrophy of tonsils (OR=4.34, 95% CI 3.72–5.07, N=309), and hypertrophy of tonsils → acute tonsillitis (OR=4.33, 95% CI 3.71–5.05, N=309). All ORs adjusted for sex, age group, calendar year, and hospital. BH-adjusted p<0.05.

Abbreviations: OR= odds ratio, CI= confidence interval, BH= Benjamini-Hochberg, N= number of patients with both diagnoses.

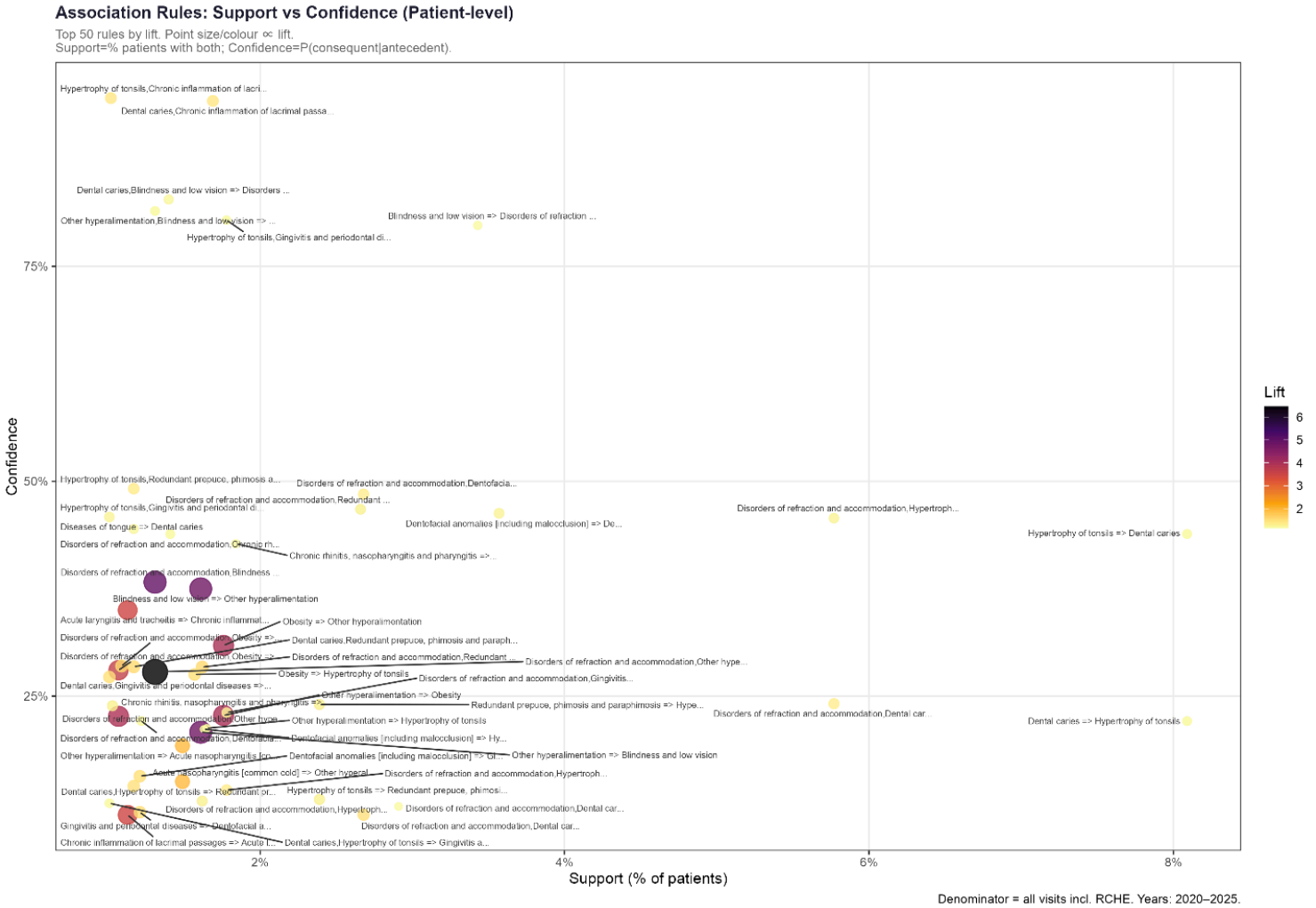


Figure 8

Association rules: support versus confidence scatter plot for the top 50 diagnostic pathways by lift.

Scatter plot showing the 50 highest-lift association rules extracted by the Apriori algorithm from patient-level diagnosis data (N=72,909 patients). Each point represents one rule of the form {antecedent} ⇒{consequent}. The x-axis shows support (proportion of patients with both antecedent and consequent diagnoses, range 1 to 8%), the y-axis shows confidence (conditional probability of the consequent given the antecedent). Point size and color both scale with lift (yellow=low, dark purple=high, range 1.2 to 6.5). Rule labels are shown for selected high-lift and high-confidence rules. A total of 110 significant rules were identified. The highest-lift rule was {disorders of refraction and accommodation, other hyperalimentation} ⇒blindness and low vision (lift=6.46, confidence=27.8%, support=1.3%). The highest-confidence rules involved disorders of refraction and accommodation as the consequent, reflecting the very high marginal prevalence of refractive disorders in this cohort. Minimum thresholds for rule extraction: support ≥0.5%, confidence ≥20%, lift ≥1.5. Denominator: all visits including RCHE. Years: 2020 to 2025.

Abbreviations: Apriori= association rule mining algorithm, RCHE= routine child health examination.

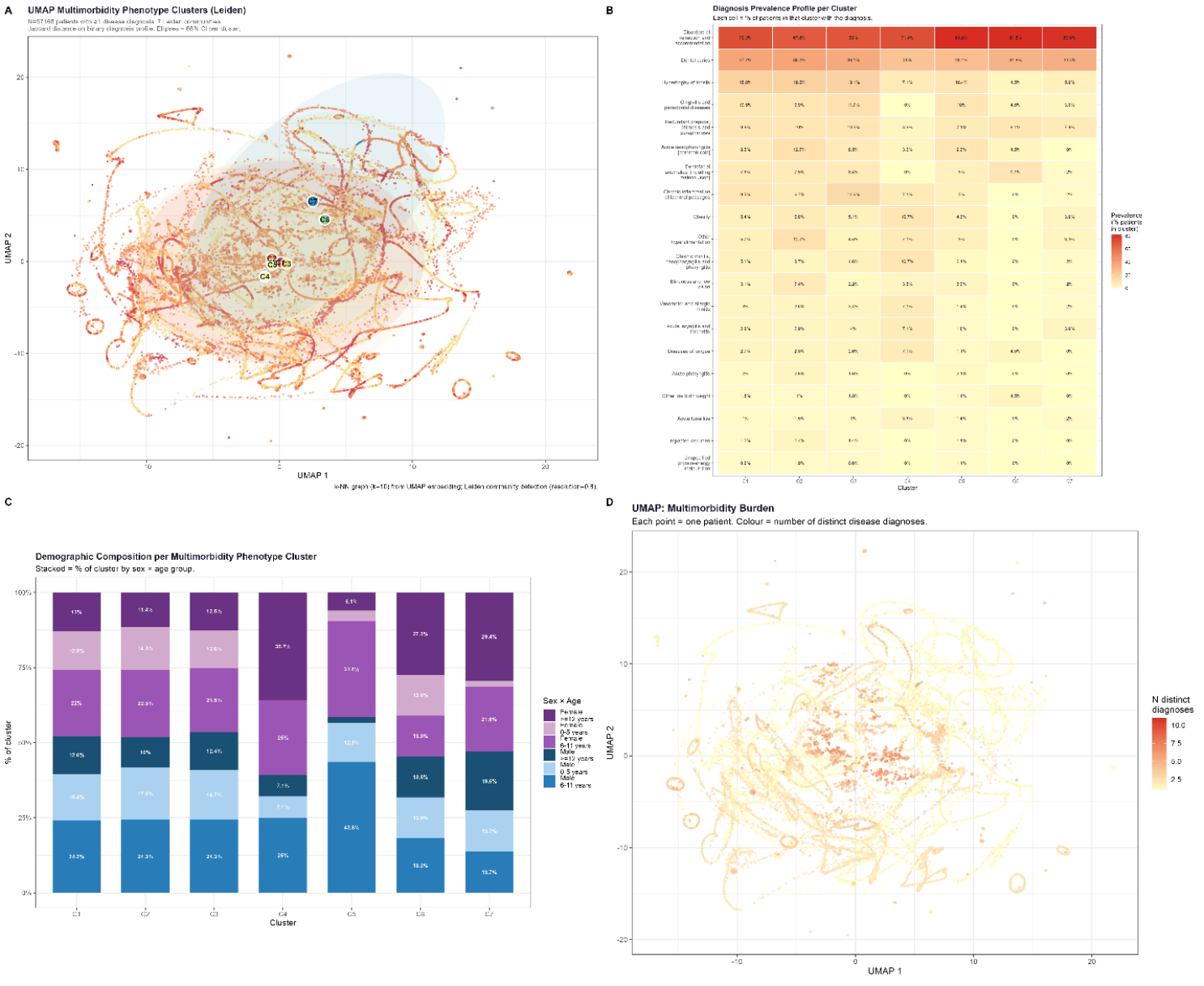


Figure 9

UMAP multimorbidity phenotype clusters: embedding, diagnosis prevalence profiles, demographic composition, and burden gradient.

Panel A. UMAP Multimorbidity Phenotype Clusters (Leiden).

Two-dimensional UMAP embedding of 57,168 patients with ≥1 disease diagnosis. Each point represents one patient, point color reflects Leiden community cluster membership (C1 to C7). UMAP was generated using Jaccard distance on binary diagnosis profiles, with parameters: n_neighbors=15, min_dist=0.1, n_epochs=500. Community detection used the Leiden algorithm with resolution=0.8 applied to a k-nearest neighbor graph (k=10) from the UMAP embedding. Ellipses represent 68% confidence intervals per cluster. Cluster centroids are labelled. Three large clusters (C1: N=10,139, C2: N=9,891, C3: N=6,166) and four additional clusters (C4 to C7) are shown.

Panel B. Diagnosis Prevalence Profile per Cluster.

Heatmap showing the percentage of patients within each cluster (columns) carrying each of the top 20 diagnoses (rows). Cell values are within-cluster prevalences (%). Color intensity reflects prevalence (white=low, red=high). Clusters were characterized by distinct enrichment patterns: C2 showed high prevalence of blindness and low vision and other hyperalimentation, defining a visual-nutritional phenotype, C3 showed enrichment for chronic inflammation of lacrimal passages, defining a lacrimal-respiratory phenotype, C4 showed strong enrichment for chronic inflammation of lacrimal passages (14.4%).

Panel C. Demographic Composition per Multimorbidity Phenotype Cluster.

Stacked bar chart showing the sex × age group composition of each cluster as a percentage of cluster membership. Stacked segments represent six sex-age combinations (Female/Male × age 0 to 5, 6 to 11, ≥12 years). The large clusters (C1 to C3) showed broadly balanced sex distributions (46.5 to 47.8% female) with similar age profiles.

Panel D. UMAP: Multimorbidity Burden.

The same UMAP embedding as Panel A, with each patient colored by number of distinct disease diagnoses (yellow=1 to 2, orange=5, red=10+). Patients with higher disease burden concentrate in the central and upper regions of the embedding space, forming a visible complexity gradient. Patients in peripheral regions predominantly carry single diagnoses. This spatial segregation suggests that multimorbidity phenotype clusters reflect underlying disease-complexity gradients not fully captured by cluster membership alone.

Abbreviations: UMAP= Uniform Manifold Approximation and Projection, k-NN= k-nearest neighbor, CI= confidence interval, C1 to C7= Leiden community clusters 1 to 7.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.xlsx
- SupplementarynoteSTROBEchecklistv4cohortcomorbidity.doc