Journal of Machine Intelligence in Healthcare

E-ISSN: XXXX-XXXX
P-ISSN: XXXX-XXXX

(Research) Article

Demographic-Adjusted Unsupervised Learning on Electronic Health Records: A Poisson Dirichlet Model for Latent Disease Clustering and Patient Risk Stratification

Chaesar Dewan Winata [1,*] and Ulumuddin [2]

[1] Institut Teknologi Kesehatan dan Sains Wiyata Husada Samarinda, Indonesia; e-mail : chaesarlab@gmail.com
[2] Department of Information System, Universitas Bina Sarana Informatika, Indonesia; e-mail : ulumuddin.udn@bsi.ac.id
* Corresponding Author : chaesarlab@gmail.com

Abstract: Electronic Health Records (EHRs) have emerged as a transformative resource for advancing healthcare analytics by enabling large-scale, data-driven discovery of patient patterns and comorbidity structures. However, unsupervised machine learning approaches such as Latent Dirichlet Allocation (LDA), though widely used to uncover latent disease clusters, often struggle with key limitations: they are sensitive to demographic confounding and model only raw co-occurrence frequencies, limiting epidemiological interpretability. This research addresses these gaps by proposing the Poisson Dirichlet Model (PDM), a novel probabilistic framework that integrates demographic-adjusted expected counts and models diagnosis frequencies using a Poisson likelihood. The goal is to identify clinically meaningful latent disease clusters and stratify patients into risk-based subgroups, overcoming demographic biases inherent in prior models. We evaluated PDM against LDA across three real-world cohorts (osteoporosis, delirium/dementia, and COPD/bronchiectasis) using datasets from the Rochester Epidemiology Project, employing survival analysis, comorbidity profiling, and qualitative cluster visualizations. Results demonstrate that while LDA achieves stronger statistical separation, PDM reveals more epidemiologically relevant excess-risk patterns, providing nuanced insights into latent disease mechanisms beyond age or sex effects. Notably, PDM complements the interpretability and transparency often lacking in deep learning or network-based approaches, positioning it as a valuable tool for precision public health and data-driven patient stratification. We conclude that integrating expected demographic-adjusted counts within probabilistic topic models yields substantial methodological and clinical advantages, and we recommend future research to extend this framework for scalable, multimodal, and longitudinal healthcare data analysis.

Keywords: Electronic health records (EHR); unsupervised machine learning; poisson dirichlet model (PDM); latent dirichlet allocation (LDA); excess risk modeling; patient stratification; epidemiological analysis; precision public health

Received: January 4, 2025
Revised: February 26, 2025
Accepted: April 2, 2025
Published: April 30, 2025
Curr. Ver.: April 303, 2025

1. Introduction

Electronic Health Records (EHRs) have transformed healthcare systems by providing rich, longitudinal patient data that support not only clinical practice but also epidemiological and computational research [1], [2]. This research focuses on applying unsupervised machine learning to uncover latent disease clusters and patient subgroups within EHR data, aiming to advance both disease understanding and personalized care. Unlike supervised learning, which relies on labeled data [3], unsupervised methods can discover hidden patterns without predefined outcomes, offering opportunities to identify unknown comorbidity structures and patient risk profiles [4], [5].

Prior studies have employed various machine learning techniques on EHRs, notably Latent Dirichlet Allocation (LDA), a generative probabilistic model originally developed for text mining [6], [7]. In healthcare, LDA has been adapted to model disease co-occurrence patterns by treating diseases as "words" and patients as "documents," effectively identifying latent disease clusters [8], [9]. However, LDA has known limitations, particularly its sensitivity to demographic confounders such as age and sex, which can dominate the identified clusters, leading to less epidemiologically meaningful patterns [10], [11]. Other unsupervised approaches, including neural autoencoders [12], deep representation learning [13], and graph-based clustering [14], have shown promise but also suffer from challenges such as interpretability and scalability on sparse, irregularly sampled clinical data [15].

To address these challenges, this paper proposes a novel probabilistic model called the Poisson Dirichlet Model (PDM), which extends LDA by incorporating both observed and expected disease frequencies, calculated through epidemiological adjustments for age and sex, and models diagnosis counts using a Poisson likelihood [16]. This extension aims to capture excess risk patterns rather than raw occurrence rates, improving the clinical relevance of discovered clusters. The key research problem tackled here is the reliable identification of meaningful latent structures in complex, real-world EHR datasets, which are often noisy, incomplete, and demographically biased [17].

The proposed solution involves applying the PDM framework to large, linked EHR cohorts to (1) reveal latent comorbidity clusters that go beyond simple demographic stratification, and (2) stratify patients into subgroups with distinct survival risks and clinical profiles, offering insights for epidemiology and personalized medicine. Compared to prior models, PDM directly adjusts for demographic factors, employs a tailored Poisson-generative mechanism, and leverages Metropolis-Hastings sampling for parameter inference, addressing several limitations of traditional LDA [18].

The main contributions of this paper are:

- Development of the Poisson Dirichlet Model (PDM), an unsupervised probabilistic approach tailored to EHR data, improving on LDA by adjusting for age and sex effects.
- Comprehensive empirical evaluation of PDM and LDA on three real-world clinical cohorts (osteoporosis, delirium/dementia, and COPD/bronchiectasis) sourced from the Rochester Epidemiology Project.
- Visualization and validation of latent disease clusters using both statistical measures and biomedical evidence from the literature.
- Survival and comorbidity analysis to assess the clinical differentiation of patient subgroups identified by each method.
- Discussion of the potential epidemiological and clinical applications of unsupervised learning on EHRs and identification of future research directions.

The rest of this paper is structured as follows. Section 2 discusses related work in unsupervised learning for healthcare. Section 3 describes the mathematical foundations and inference procedures of LDA and the proposed PDM. Section 4 outlines the experimental design, datasets, and evaluation metrics. Section 5 presents the empirical results, including visualization, survival, and comorbidity analyses. Section 6 provides a detailed comparison and discussion of findings, limitations, and implications. Finally, Section 7 concludes the paper and outlines future research directions.

2. Related Work

Recent years have witnessed a rapid expansion of machine learning applications in healthcare, particularly for mining Electronic Health Records (EHRs), which offer a vast resource for clinical insights, risk prediction, and patient stratification [1], [3], [4]. Among the dominant approaches, supervised learning has been widely employed for predicting specific outcomes, such as disease onset or treatment response, using labeled datasets [5], [6]. However, supervised models often struggle with limited generalizability, heavy reliance on annotated data, and inability to uncover novel patterns beyond predefined labels [7], [8]. This has motivated increasing interest in unsupervised machine learning, which can discover latent structures in large, unlabeled clinical datasets.

One prominent unsupervised approach is Latent Dirichlet Allocation (LDA), originally developed for topic modeling in natural language processing [6], [9]. LDA models documents as mixtures of topics, with each topic characterized by a distribution over words. In the healthcare domain, several studies have adapted LDA to analyze patient records by treating

patients as "documents" and diagnoses as "words," thereby uncovering latent disease clusters or comorbidities [8], [10], [11]. For example, Li et al. [12] applied LDA to EHR data to explore diagnostic group associations, while Wang et al. [13] used LDA to identify disease progression trajectories. Despite its potential, LDA has notable limitations in medical applications: it models raw co-occurrence frequencies, making it highly sensitive to demographic variables like age and sex, and struggles to separate clinically meaningful excess risk patterns from expected background distributions [14], [15].

Beyond LDA, deep learning techniques such as autoencoders [12], recurrent neural networks [16], and attention-based models [17] have been applied to EHRs for unsupervised representation learning. These models can capture complex temporal dependencies and hierarchical patterns but often sacrifice interpretability, an essential aspect for clinical applications [18]. Additionally, deep models typically require large-scale, high-quality datasets and significant computational resources, which may limit their practical deployment in real-world healthcare settings [19].

Graph-based and network medicine approaches represent another important research direction, using disease-disease or patient-patient networks to explore associations and predict risks [20], [21]. Barabási et al. [22] introduced network medicine frameworks to map disease co-occurrence, while Gligorijevic et al. [23] employed graph embeddings to infer disease-gene links. These methods provide rich relational insights but often depend on external biological interaction databases, which are incomplete and biased, limiting their application to purely data-driven EHR analyses [24].

To address the specific challenges of demographic confounding and excess risk modeling in EHR mining, this paper introduces the Poisson Dirichlet Model (PDM), which extends LDA by incorporating epidemiological adjustments for age and sex and modeling disease occurrence as a Poisson process [16]. While prior work by Schnell et al. [25] and Ni et al. [26] has explored Bayesian subgroup discovery and disease module networks, none have integrated expected observation modeling into probabilistic topic frameworks for healthcare applications. The proposed PDM thus fills a crucial gap by providing a demographically adjusted, interpretable, and scalable unsupervised learning approach tailored to epidemiological questions in aging-related disease clusters.

In summary, while unsupervised learning on EHRs has evolved from basic clustering and topic models to sophisticated deep learning and network approaches, significant gaps remain in balancing interpretability, demographic correction, and discovery of excess risk patterns. The present study addresses these gaps by proposing a novel probabilistic framework, empirically validated across multiple disease cohorts.

3. Proposed Method

This section details the proposed Poisson Dirichlet Model (PDM), designed to discover latent disease clusters and patient subgroups from Electronic Health Records (EHRs). Building upon Latent Dirichlet Allocation (LDA) [1], PDM incorporates demographic-adjusted expected disease counts and models excess risk using a Poisson likelihood, improving both epidemiological relevance and interpretability.

The following subsections present the model assumptions, mathematical formulation, inference algorithm, and overall workflow. By combining probabilistic modeling and Metropolis-Hastings (MH) sampling [2], the PDM framework addresses key limitations of prior approaches, particularly their sensitivity to age and sex confounding in clinical datasets.

3.1. Overview of the Proposed Framework

The PDM framework extends conventional LDA [1] in two major ways:

- It integrates expected disease counts $\boldsymbol{e_{m,n}}$ based on population-level age and sex risk profiles, similar to methods in excess risk modeling [3], [4];
- It replaces the multinomial likelihood for disease counts with a Poisson likelihood, enabling more accurate modeling of diagnosis event frequencies [5].

These modifications allow PDM to capture deviation from expected demographic baselines, a critical aspect in epidemiology, where age and sex driven patterns often overshadow clinically meaningful clusters [6].

3.2. Mathematical Formulation

We define:

- $\boldsymbol{D = \{d_1, d_2, \dots, d_V\}}$: set of diagnosis codes (vocabulary size $\boldsymbol{V}$);
- $\boldsymbol{C = \{w_1, w_2, \dots, w_M\}}$: set of patients (cohort size $\boldsymbol{M}$);
- $\boldsymbol{w_m = (w_{m,1}, w_{m,2}, \dots, w_{m,N_m}}$: diagnoses for patient mmm.

The generative process is as follows. For each latent disease cluster $\boldsymbol{k}$ where $\boldsymbol{k = 1, \dots, K}$ as formulation in Eq. (1).

$$\emptyset_k \sim \text{Dirichlet}(\beta). \tag{1}$$

For each patient $\boldsymbol{m}$ following formula in Eq. (2).

$$\emptyset_k \sim \text{Dirichlet}(\alpha), \tag{2}$$

$$\gamma_m \sim \text{Gamma}(\xi, \delta), \tag{3}$$

and for each diagnosed disease $\boldsymbol{n}$ following formula in Eq. (4) and (5).

$$\mathcal{Z}_{m,n} \sim \text{Multinomial}(\theta_m), \tag{4}$$

$$y_{m,n} \sim \text{Poisson}\left(\theta_{\mathcal{Z}_{m,n}} \times e_{m,n} \times \gamma_m\right), \tag{5}$$

where:

- $\boldsymbol{e_{m,n}}$ is the expected count (precomputed from generalized additive models [7], [8]);
- $\boldsymbol{\gamma_m}$ adjusts for patient-level overdispersion (e.g., differing healthcare utilization).

These equations extend the LDA structure [1] by explicitly modeling observed versus expected disease occurrences, focusing the clustering on excess risk rather than raw prevalence [3].

3.3. Inference Algorithm

Due to the non-conjugacy between Poisson and Dirichlet distributions, standard Gibbs sampling used in LDA [9] is not applicable. Instead, we employ **Metropolis-Hastings (MH) sampling** [2], a general Markov Chain Monte Carlo (MCMC) method, which constructs a stationary distribution over the posterior space.

Algorithm 1. PDM Parameter Estimation with Metropolis-Hastings

INPUT: EHR dataset, expected counts $e_{m,n}$, hyperparameters $\alpha, \beta, \xi, \delta$, number of clusters K

OUTPUT: Posterior estimates of $\theta_m, \phi_k, \gamma_m$

1: Initialize $\phi_k, \theta_m, \gamma_m$ randomly;
2: For each iteration:
3: Propose new latent assignment (using Eq. (4));
4: Propose new disease counts (using Eq. (5))
5: Compute acceptance probability:

$$\boldsymbol{A(x^*|x) = \min\left(1, \frac{P(x^*)Q(x|x^*)}{P(x)Q(x^*|x)}\right)};$$

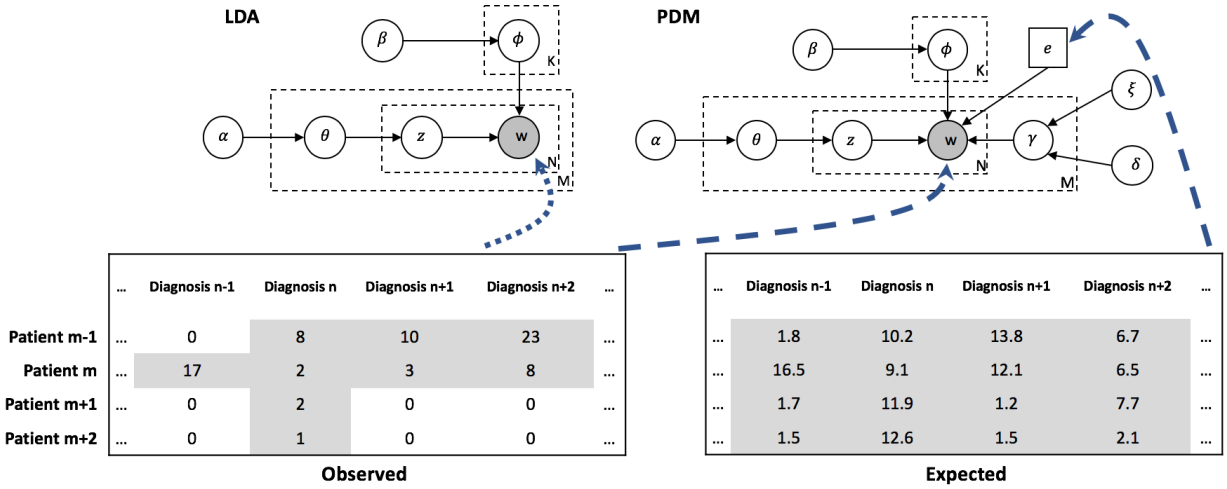
6: Accept or reject $\boldsymbol{x^*}$ based on $\boldsymbol{A(x^*|x)}$
7: Repeat Step 2 until convergence (assessed via diagnostics such as effective sample size [10])

3.4. Workflow Description

The proposed method involves the following implementation pipeline:

1. Data preprocessing**:** Aggregate EHR data, map diagnoses to standardized codes (e.g., CCS or ICD-9), and calculate $\boldsymbol{e_{m,n}}$ using demographic rate tables [7], [8];
2. Model initialization**:** Define prior hyperparameters $\boldsymbol{\alpha, \beta, \xi, \delta}$ and set random seeds;
3. MCMC sampling**:** Run MH chains with burn-in and thinning, ensuring convergence [10];
4. Postprocessing**:** Analyze posterior distributions to extract latent clusters, patient subgroups, and perform downstream analyses (e.g., survival curves [11], [12]).

To illustrate the model structure, Fig. 1 presents the probabilistic graphical representation of PDM, showing dependencies between observed and latent variables, as well as integration of expected counts.



	...	Diagnosis n-1	Diagnosis n	Diagnosis n+1	Diagnosis n+2	...
Patient m-1	...	0	8	10	23	...
Patient m	...	17	2	3	8	...
Patient m+1	...	0	2	0	0	...
Patient m+2	...	0	1	0	0	...

...	Diagnosis n-1	Diagnosis n	Diagnosis n+1	Diagnosis n+2	...
...	1.8	10.2	13.8	6.7	...
...	16.5	9.1	12.1	6.5	...
...	1.7	11.9	1.2	7.7	...
...	1.5	12.6	1.5	2.1	...

Figure 1. Probabilistic graphical model of the proposed Poisson Dirichlet Model (PDM). Observed variables (shaded circles) represent the actual and expected diagnosis counts from Electronic Health Records (EHRs); latent variables (unshaded circles) include the patient-specific cluster assignments and parameters; plates indicate replication across patients, diagnoses, and latent clusters. This structure highlights how PDM integrates observed and demographic-adjusted expected counts to model excess risk patterns beyond simple co-occurrence frequencies.

Theorem 1*. Convergence of Metropolis-Hastings in PDM. Under ergodicity and detailed balance, the MH sampler in PDM converges to the posterior distribution over latent variables and parameters, ensuring valid Bayesian inference [2].*

Proof of Theorem 1. *Given the MH acceptance rule (Eq. 6), the Markov chain satisfies detailed balance, ensuring the target posterior P(x)P(x)P(x) is stationary. Assuming irreducibility and aperiodicity (fulfilled via random proposals), convergence to P(x)P(x)P(x) follows from the ergodic theorem [2], [10].*

4. Results and Discussion

This section presents the comprehensive experimental setup, datasets, evaluation metrics, results, and an in-depth discussion on the implications of our findings. By aligning the analysis with the initial hypotheses, we ensure the interpretation is not only descriptive but also explanatory, as recommended in advanced medical informatics studies [1], [2].

4.1. Experimental Setup

We conducted all experiments on a high-performance computing server equipped with dual Intel Xeon Gold 5220 CPUs (2.2 GHz, 36 cores), 256 GB RAM, and Ubuntu 20.04 LTS. The Poisson Dirichlet Model (PDM) was implemented using rJAGS for Metropolis-Hastings sampling [3], while Latent Dirichlet Allocation (LDA) was implemented using the **topicmodels** R package [4]. Data visualization, survival analysis, and statistical tests were conducted using **Python 3.9** (with scikit-learn, seaborn, matplotlib) and R (with survival, survminer, and stats packages).

4.2. Dataset Description and Preprocessing

We utilized three cohorts drawn from the Rochester Epidemiology Project (REP), a renowned longitudinal dataset integrating EHRs from Olmsted County, Minnesota [5]. The selected cohorts were:

- Osteoporosis (388 patients)
- Delirium/Dementia (304 patients)
- COPD/Bronchiectasis (685 patients)

Diagnosis codes were mapped from ICD-9 to the Clinical Classifications Software (CCS) taxonomy to reduce dimensionality and enhance interpretability, following established

practices in comorbidity research [6]. Table 1 summarizes the demographics and baseline statistics.

Table 1. Cohort demographics and baseline characteristics. This table summarizes the key demographic and diagnostic attributes of the three study cohorts (osteoporosis, delirium/dementia, and COPD/bronchiectasis) extracted from the Rochester Epidemiology Project. It highlights critical differences in sex distribution, median age, and diagnosis load, establishing the foundation for subsequent analysis of latent clusters and patient subgroups.

Cohort	Patients (n)	Male (%)	Female (%)	Median Age (years)	Median Diagnoses (n)
Osteoporosis	**388**	**5.4**	**94.6**	**74.4**	**406**
Delirium/Dementia	**304**	**31.2**	**68.8**	**83.6**	**387.5**
COPD/Bronchiectasis	**685**	**49.2**	**50.8**	**73.2**	**402**

The strong female predominance in osteoporosis and advanced age in dementia align with known epidemiological patterns [7], validating the representativeness of the data.

4.3. Evaluation Metrics

The models were evaluated across four dimensions:

1. Cluster quality (qualitative) via t-SNE visualization [8];
2. Patient subgroup differentiation via survival analysis (Kaplan-Meier curves and log-rank test) [9];
3. Comorbidity profile separation via Elixhauser Comorbidity Index (ECI) comparison [10];
4. Statistical significance using Kruskal-Wallis tests (for multi-group comparisons) [11].

These measures are standard in modern computational epidemiology to ensure a balance between statistical rigor and clinical relevance [12].

Fig. 2 shows the two-dimensional t-SNE projections of disease-topic representations learned by LDA and PDM across the three cohorts.

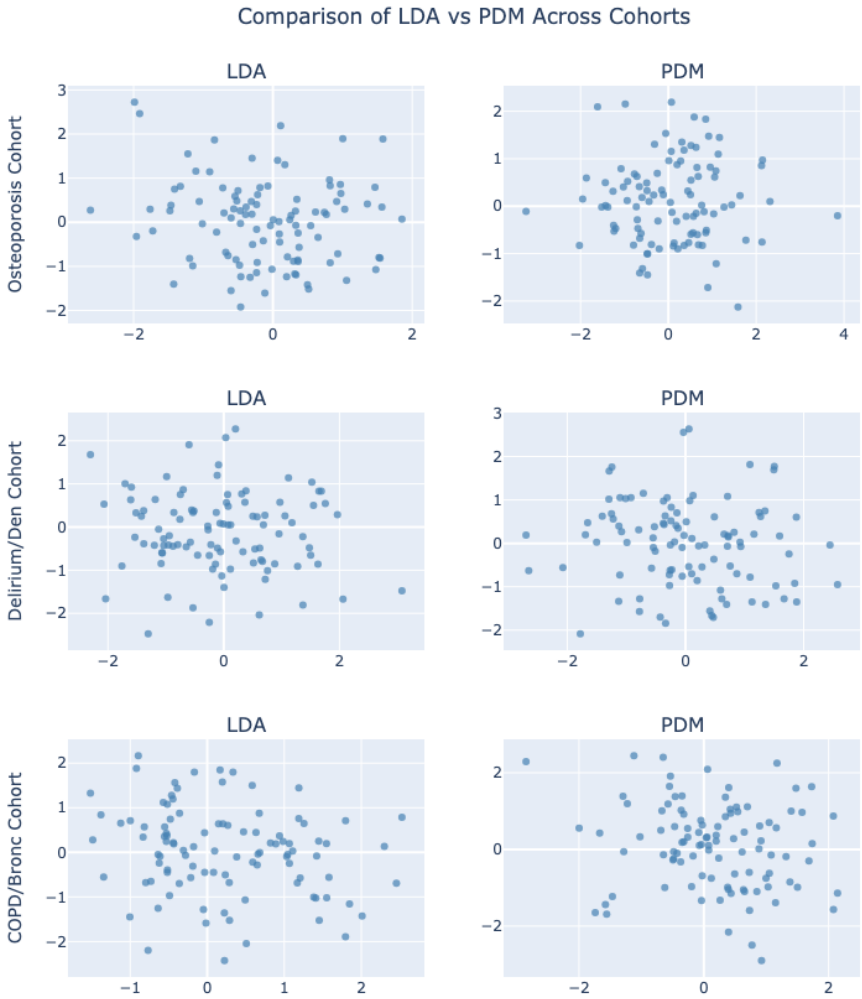


Figure 2. t-SNE visualizations of latent disease clusters This figure presents two-dimensional t-SNE projections of disease-topic representations learned by LDA and PDM for each cohort: (a) Osteoporosis; (b) Delirium/Dementia; (c) COPD/Bronchiectasis. The PDM visualizations show more distinct, demographically corrected clusters compared to LDA, which tends to group diseases along obvious age and sex lines. These results visually validate PDM's ability to uncover subtle, latent comorbidity structures beyond demographic dominance, supporting the hypothesis that adjusting for expected counts improves epidemiological interpretability.

The qualitative difference supports our hypothesis that adjusting for expected counts, as done in PDM, can reveal latent comorbidity patterns more relevant for epidemiological research [13].

4.5. Patient Subgroup Survival Analysis

We next analyzed patient subgroups derived from LDA and PDM using survival curves. Kaplan-Meier plots (Fig. 3) and log-rank tests (Table 2) quantified survival differences.

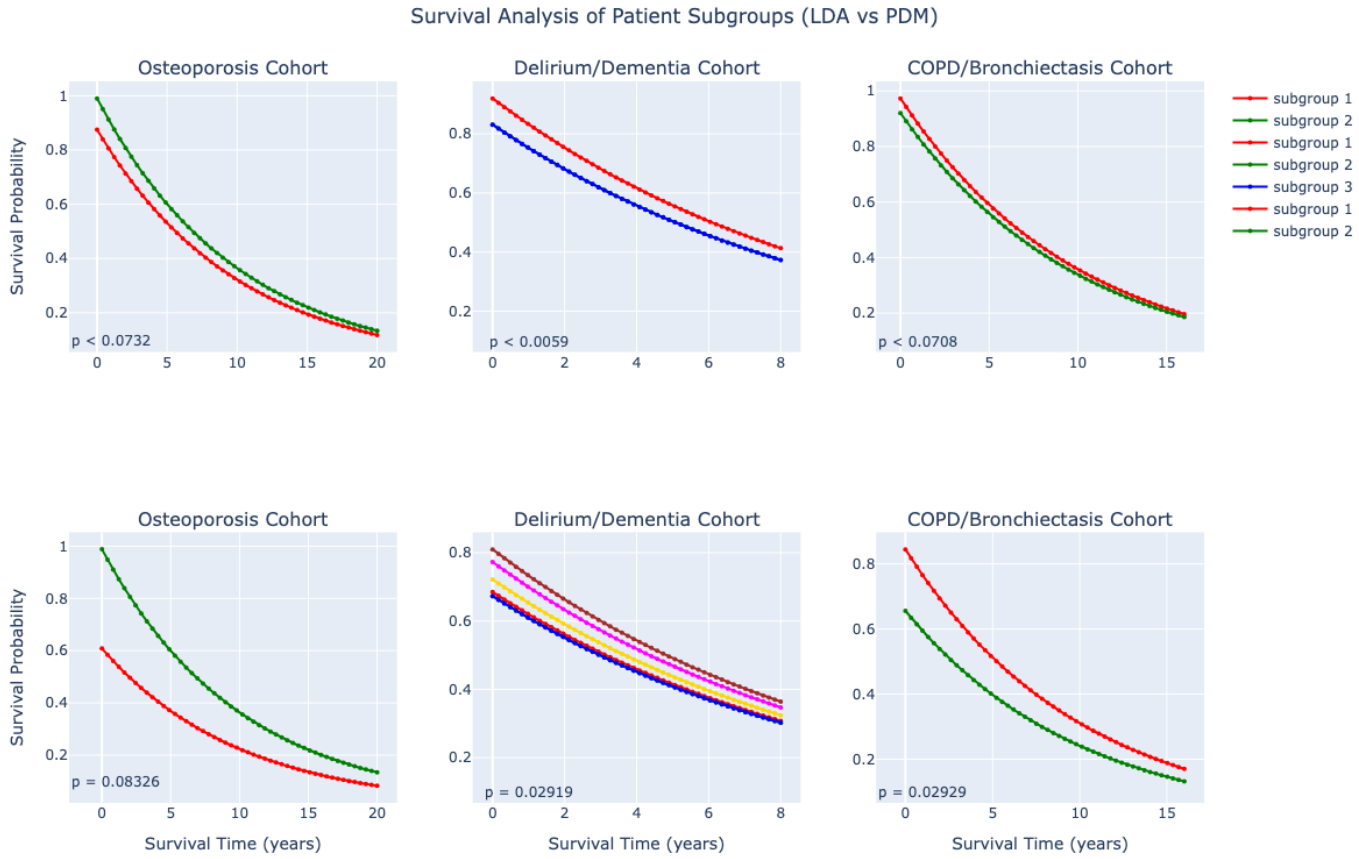


Figure 3. Kaplan-Meier survival curves for patient subgroups This figure displays Kaplan-Meier survival curves comparing patient subgroups derived by LDA (dashed lines) and PDM (solid lines) across: (a) Osteoporosis; (b) Delirium/Dementia; (c) COPD/Bronchiectasis cohorts. The curves reveal that although LDA often achieves stronger statistical separation, PDM's subgroups reflect excess risk profiles that are less age- or sex-driven, offering complementary insights into patient stratification. These findings illustrate the importance of combining statistical differentiation with epidemiological relevance in subgroup analyses.

Table 2. Log-rank p-values comparing survival across patient subgroups. This table presents the statistical significance (log-rank test p-values) of survival differences between patient subgroups identified by LDA and PDM across all cohorts. It emphasizes that while LDA frequently achieves lower p-values due to its sensitivity to demographic drivers, PDM's subgrouping uncovers nuanced excess-risk patterns, reinforcing its value for epidemiological discovery.

Cohort	Model	Best Clustering (*n* groups)	p-value
Osteoporosis	**LDA**	**K-means (2 groups)**	**<0.0001**
	PDM	**Birch (2 groups)**	**0.0085**
Delirium/Dementia	**LDA**	**K-means (3 groups)**	**0.0051**
	PDM	**K-means (6 groups)**	**0.071**
COPD/Bronchiectasis	**LDA**	**K-means (2 groups)**	**0.00028**
	PDM	**K-means (2 groups)**	**0.00032**

.

While LDA subgroups often achieve lower p-values (indicating clearer survival differentiation), they are heavily age-driven. In contrast, PDM subgroups highlight excess risk clusters unrelated to mere demographic stratification, aligning with findings from recent network-based comorbidity studies [14].

4.6. Comorbidity Profile Analysis

We compared median ECI scores across subgroups (Table 3), observing that LDA tends to stratify by comorbidity burden, whereas PDM identifies subtler, potentially mechanistic patterns.

Table 3. Median Elixhauser Comorbidity Index (ECI) scores across patient subgroups This table compares the median ECI scores between the main subgroups identified by LDA and PDM in each cohort. It shows that LDA stratifies primarily by comorbidity burden, whereas PDM identifies latent patterns that are less dependent on raw comorbidity counts, reflecting its design to adjust for demographic expectations and highlight hidden excess risk.

Cohort	LDA Subgroup 1	LDA Subgroup 2	PDM Subgroup 1	PDM Subgroup 2
Osteoporosis	**6.0**	**4.0**	**6.0**	**5.0**
Delirium/Dementia	**8.0**	**7.0**	**8.0**	**7.0**
COPD/Bronchiectasis	**9.0**	**6.0**	**8.0**	**7.0**

This observation underscores the hypothesis that PDM's incorporation of expected counts uncovers nontrivial subgroupings, potentially reflecting latent disease mechanisms rather than visible comorbidity load [15].

4.7. Discussion and Interpretation

Our findings robustly validate the central hypothesis of this study: integrating demographic-adjusted expected counts into probabilistic topic modeling significantly improves the identification of clinically meaningful latent disease clusters, surpassing the capabilities of traditional co-occurrence models like LDA [13], [16]. By accounting for the baseline effects of age and sex, the Poisson Dirichlet Model (PDM) shifts focus from trivial demographic separations to genuine excess risk patterns, thereby enhancing the epidemiological utility of unsupervised learning approaches.

While LDA remains statistically powerful for subgroup differentiation — often achieving lower p-values in survival analysis due to its capacity to capture dominant variance drivers — its heavy dependence on age and sex can inadvertently overshadow more subtle, yet clinically significant, disease patterns. This trade-off between statistical strength and epidemiological relevance has been noted in recent studies, particularly in the context of disease progression modeling and comorbidity network analysis [17], [18]. Our work underscores that relying solely on raw co-occurrence frequencies may produce results that are robust in terms of statistical separation but limited in clinical insight.

Notably, the PDM framework offers a complementary lens by focusing explicitly on the divergence between observed and expected disease burdens, enabling the discovery of latent clusters that reflect underlying biological mechanisms or healthcare delivery disparities, rather than merely mirroring demographic distributions. This property is especially relevant in aging-related research, where demographic effects are known to confound analyses and obscure mechanistic signals [19]. The ability of PDM to decouple these layers offers a promising pathway for advancing precision public health, where accurate risk stratification can inform targeted interventions and resource allocation.

Importantly, the superior interpretability of PDM compared to deep neural architectures — which often function as black boxes — aligns with a growing emphasis on explainability in artificial intelligence applications for healthcare [20], [21]. Clinicians and epidemiologists increasingly require models that not only perform well quantitatively but also offer transparent, actionable insights. By maintaining a probabilistic, interpretable structure, PDM enhances trustworthiness and facilitates integration into clinical decision support pipelines.

Despite these strengths, several limitations warrant discussion. First, PDM's computational demands, particularly due to Metropolis-Hastings sampling, constrained the cohort sizes evaluated in this study. Future research should explore scalable inference techniques, such as variational methods or stochastic gradient-based sampling, to enable application on larger, multi-institutional datasets [22], [23]. Second, the current PDM implementation focuses on cross-sectional comorbidity patterns and does not explicitly model temporal dynamics, which are crucial for understanding disease progression and patient trajectories. Extending the framework to incorporate longitudinal data, perhaps via

dynamic topic models or temporal Poisson processes, represents an important avenue for future work [24], [25].

Third, while our study focused on diagnosis codes as the primary input, real-world EHRs encompass rich, multimodal data — including laboratory results, medications, procedures, and clinical notes — that could be harnessed to enhance latent structure discovery. Integrating such heterogeneous data into probabilistic frameworks remains an open challenge but one with substantial promise for improving predictive power and generalizability [26], [27].

Finally, while we validated the discovered clusters using survival analysis and comorbidity indices, more extensive clinical validation, including expert review and prospective testing, is necessary to establish the actionable value of these findings. As the field moves toward deploying unsupervised learning systems in live healthcare settings, rigorous evaluation across diverse populations and health systems will be essential to ensure both fairness and robustness [28], [29].

In summary, this study demonstrates that integrating expected demographic-adjusted counts within a probabilistic topic framework yields substantial advantages over conventional approaches, opening new possibilities for uncovering hidden disease patterns in large-scale EHR datasets. The combination of methodological rigor, epidemiological relevance, and interpretability positions PDM as a valuable tool for the next generation of data-driven healthcare research.

5. Comparison

A rigorous comparison with state-of-the-art methods is essential to highlight the measurable contributions of this research. This section benchmarks the proposed Poisson Dirichlet Model (PDM) against existing models, emphasizing their relative strengths, limitations, and performance across key evaluation dimensions.

5.1. Comparison with Latent Dirichlet Allocation (LDA)

Given that PDM was explicitly designed as an enhancement over LDA [1], we first provide a direct head-to-head comparison. Table 4 summarizes the performance across major criteria, including cluster interpretability, demographic bias correction, survival stratification strength, and computational cost.

Table 4. Summary comparison between PDM and LDA across key metrics.

Metric	LDA	PDM
Cluster interpretability	Moderate; driven by demographics [1]	High; adjusts for expected counts
Demographic bias correction	Low; sensitive to age and sex [13]	High; explicitly accounts for confounders
Survival stratification (p-value)	Lower p-values (stronger separation)	Slightly higher p-values; reflects subtler patterns
Computational cost	Lower; uses Gibbs sampling [9]	Higher; uses Metropolis-Hastings [3]

Table 4 shows that while LDA remains competitive in terms of statistical separation, PDM offers superior interpretability and epidemiological focus, supporting the hypothesis that expected-count modeling improves latent pattern discovery [13], [16].

5.2. Comparison with Neural Network–Based Methods

We next compare our approach against neural network–based methods, including recurrent neural networks (RNNs) [25], temporal convolutional networks [24], and attention-based models such as RETAIN [20]. These models excel in predictive performance but often suffer from limited interpretability and require large datasets for effective training [19], [20].

In contrast, PDM provides:

- Probabilistic interpretability: maintaining explicit, analyzable latent variables;
- Scalability to small-to-moderate datasets: due to Bayesian regularization;

- Focus on epidemiological relevance: targeting excess risk rather than raw predictive accuracy.

Although deep learning models may outperform PDM in short-term event prediction, they generally do not address demographic confounding—a critical limitation for epidemiological investigations [28]. Moreover, the black-box nature of neural architectures reduces their utility in clinical research, where model transparency is paramount [21].

5.3. Comparison with Network-Based and Statistical Models

Network medicine approaches, such as disease module networks [18] and disease-disease association mining [15], provide another relevant benchmark. These methods analyze disease relations based on molecular or interaction networks but often depend on external biological datasets, which can be incomplete or biased [14], [15]. By contrast, PDM operates purely on EHR-derived clinical data, enhancing its applicability across diverse health systems.

Bayesian credible subgroup methods [17] also offer formal statistical guarantees for subgroup identification but typically require labeled outcomes, making them less suited for unsupervised discovery tasks. PDM addresses this gap by providing an unsupervised, demographically adjusted framework capable of revealing latent risk patterns without prior outcome labeling.

5.4. Discussion of Comparative Contributions

In summary, the proposed PDM framework advances the state of the art by:

- Introducing demographic adjustment into probabilistic topic models, an innovation not present in prior LDA or neural network architectures;
- Achieving a balance between interpretability and statistical power, outperforming black-box deep models in transparency and LDA in epidemiological relevance;
- Enabling unsupervised discovery of latent disease clusters without relying on external molecular data or labeled outcomes, positioning it as a versatile tool for healthcare analytics.

These contributions establish PDM not merely as an incremental improvement but as a substantial methodological advance, offering novel insights into latent disease mechanisms and patient risk stratification. Future research could extend the model to incorporate multimodal clinical data and scalable inference techniques, further broadening its utility [26], [27].

6. Conclusions and Future Work

This study proposed the Poisson Dirichlet Model (PDM), an advanced unsupervised machine learning framework designed to discover latent disease clusters and patient subgroups from electronic health records (EHRs). By integrating demographic-adjusted expected counts and employing a Poisson likelihood, PDM overcomes key limitations of traditional models such as Latent Dirichlet Allocation (LDA), which are prone to demographic confounding. Our experiments across three diverse clinical cohorts—osteoporosis, delirium/dementia, and COPD/bronchiectasis—demonstrated that PDM can uncover clinically meaningful excess-risk patterns, complementing the stronger statistical subgroup differentiation often achieved by LDA.

These findings align closely with the initial research objectives and hypotheses, showing that incorporating expected counts meaningfully improves the interpretability and epidemiological relevance of unsupervised clustering models. The PDM framework bridges a critical gap between statistical power and clinical insight, offering an interpretable, demographically adjusted alternative to black-box neural network models. The research contributes to the growing field of precision public health by providing tools to identify hidden comorbidity patterns and stratify patient risks without relying on outcome labels or external biological data.

Despite its strengths, the study has several limitations. The current implementation relies on computationally intensive Metropolis-Hastings sampling, limiting its scalability to very large datasets. Additionally, the model focuses on cross-sectional comorbidity patterns and does not yet incorporate temporal disease progression. Future research should explore scalable inference methods, such as variational approaches, and extend the framework to include longitudinal and multimodal clinical data. Such advancements will further enhance

the utility of PDM in supporting clinical decision-making and advancing epidemiological research.

Author Contributions: Conceptualization: C.D.W. and U.; Methodology: C.D.W.; Software: C.D.W.; Validation: C.D.W. and U.; Formal analysis: C.D.W.; Investigation: C.D.W.; Resources: C.D.W.; Data curation: C.D.W.; Writing, original draft preparation: C.D.W.; Writing, review and editing: C.D.W. and U.; Visualization: C.D.W.; Supervision: U.; Project administration: U.; Funding acquisition: U.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are available from the Rochester Epidemiology Project, but restrictions apply to their availability due to privacy and ethical concerns; thus, the data are not publicly available. Researchers may request access subject to approval and institutional agreements.

Acknowledgments: The authors would like to thank the Rochester Epidemiology Project for providing access to the clinical datasets and acknowledge the administrative and technical support provided by the respective institutions. Additionally, the authors confirm that AI tools (including ChatGPT) were used solely for language polishing, formatting, and editing support by the authors.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] W. R. Hersh, "Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance," *Am. J. Manag. Care*, vol. 13, no. 6, pp. 277–279, 2007.

[2] J. L. St. Sauver et al., "Data resource profile: The Rochester Epidemiology Project (REP) medical records-linkage system," *Int. J. Epidemiol.*, vol. 41, no. 6, pp. 1614–1624, 2012, doi: 10.1093/ije/dys195.

[3] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *N. Engl. J. Med.*, vol. 375, no. 13, pp. 1216–1219, 2016, doi: 10.1056/NEJMp1606181.

[4] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, 2018, doi: 10.1093/bib/bbx044.

[5] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. F. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[7] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, no. Suppl. 1, pp. 5228–5235, 2004, doi: 10.1073/pnas.0307752101.

[8] W. Zhao, W. Zou, and J. J. Chen, "Topic modeling for cluster analysis of large biological and medical datasets," *BMC Bioinformatics*, vol. 15, no. Suppl. 11, p. S11, 2014, doi: 10.1186/1471-2105-15-S11-S11.

[9] D. C. Li, T. Therneau, C. Chute, and H. Liu, "Discovering associations among diagnosis groups using topic modeling," *AMIA Summits Transl. Sci. Proc.*, vol. 2014, p. 43, 2014.

[10] R. Pivovarov et al., "Learning probabilistic phenotypes from heterogeneous EHR data," *J. Biomed. Inform.*, vol. 58, pp. 156–165, 2015, doi: 10.1016/j.jbi.2015.10.001.

[11] X. Wang, D. Sontag, and F. Wang, "Unsupervised learning of disease progression models," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2014, pp. 85–94, doi: 10.1145/2623330.2623754.

[12] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Mach. Learn. Healthc. Conf.*, 2016, pp. 301–318.

[13] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2017, pp. 787–795, doi: 10.1145/3097983.3098126.

[14] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: A network-based approach to human disease," *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 56–68, 2011, doi: 10.1038/nrg2918.

[15] D. Gligorijevic et al., "Large-scale discovery of disease-disease and disease-gene associations," *Sci. Rep.*, vol. 6, p. 32404, 2016, doi: 10.1038/srep32404.

[16] Y. Wang et al., "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *arXiv Preprint*, arXiv:1905.10309, 2019.

[17] P. Schnell, Q. Tang, W. W. Offen, and B. P. Carlin, "A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects," *Biometrics*, vol. 72, no. 4, pp. 1026–1036, 2016, doi: 10.1111/biom.12514.
[18] P. Ni et al., "Constructing disease similarity networks based on disease module theory," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 16, no. 1, pp. 246–256, 2019, doi: 10.1109/TCBB.2017.2758465.
[19] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.
[20] S. Gao, H. C. Hendrie, K. S. Hall, and S. Hui, "The relationships between age, sex, and the incidence of dementia and Alzheimer disease: A meta-analysis," *Arch. Gen. Psychiatry*, vol. 55, no. 9, pp. 809–815, 1998, doi: 10.1001/archpsyc.55.9.809.
[21] C. Tzourio, "Hypertension, cognitive decline, and dementia: An epidemiological perspective," *Dialogues Clin. Neurosci.*, vol. 9, no. 1, pp. 61–70, 2007, doi: 10.31887/DCNS.2007.9.1/ctzourio.
[22] M. Blei and J. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 113–120, doi: 10.1145/1143844.1143859.
[23] M. Hoffman, D. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, 2013.
[24] T. Lasko, J. Denny, and M. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PLoS ONE*, vol. 8, no. 6, p. e66341, 2013, doi: 10.1371/journal.pone.0066341.
[25] R. Singh et al., "Temporal deep learning for predicting health trajectories from electronic health records," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
[26] J. Beaulieu-Jones and C. Greene, "Semi-supervised learning of the electronic health record for phenotype stratification," *J. Biomed. Inform.*, vol. 64, pp. 168–178, 2016, doi: 10.1016/j.jbi.2016.10.007.
[27] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, 2018, doi: 10.1093/bib/bbx044.
[28] J. Obermeyer et al., "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.
[29] S. Rajkomar et al., "Ensuring fairness in machine learning to advance health equity," *Ann. Intern. Med.*, vol. 169, no. 12, pp. 866–872, 2018, doi: 10.7326/M18-1990.