Quantifying Emergency Medicine Residency Learning Curves Using Natural Language Processing: A Retrospective Cohort Study

Carl Preiksaitis, Joshua Hughes, Rana Kabeer, William Dixon, Christian Rose

Submitted to: JMIR Medical Education
on: August 13, 2025

Table of Contents

Quantifying Emergency Medicine Residency Learning Curves Using Natural Language Processing: A Retrospective Cohort Study

Carl Preiksaitis[1] MD, MEd; Joshua Hughes[1] MD; Rana Kabeer[1] MD, MPH; William Dixon[1] MD, MSED; Christian Rose[1] MD

[1] Department of Emergency Medicine, Stanford University School of Medicine Palo Alto US

Corresponding Author:
Carl Preiksaitis MD, MEd

Department of Emergency Medicine, Stanford University School of Medicine
900 Welch Road
Suite 350
Palo Alto
US

Abstract

Background: The optimal duration of emergency medicine (EM) residency training remains a subject of national debate, with the Accreditation Council for Graduate Medical Education considering standardizing all programs to four years. However, empirical data on how residents accumulate clinical exposure over time are limited. Traditional measures, such as case logs and diagnostic codes, often fail to capture the breadth and depth of diagnostic reasoning. Natural language processing (NLP) of clinical documentation offers a novel approach to quantify clinical experiences more comprehensively.

Objective: This study aimed to: (1) quantify how EM residents acquire clinical topic exposure over the course of training; (2) evaluate variation in exposure patterns across residents and classes; and (3) assess changes in workload and case complexity over time to inform the discussion on optimal program length.

Methods: We conducted a retrospective cohort study of EM residents at Stanford Hospital, analyzing 244,255 emergency department encounters from July 1, 2016, to November 30, 2023. The sample included 62 residents across four graduating classes (2020–2023), representing all primary training site encounters where residents served as primary or supervisory providers. Using a retrieval-augmented generation NLP pipeline, we mapped resident clinical documentation to the 895 subcategories of the 2022 Model for Clinical Practice of Emergency Medicine (MCPEM) via intermediate mapping to the SNOMED CT CORE Problem List Subset. We generated cumulative topic exposure curves, quantified the diversity of topic coverage, assessed variability between residents, and analyzed progression in clinical complexity using Emergency Severity Index (ESI) scores and admission rates.

Results: Residents encountered the largest increase in new topics during postgraduate year 1 (PGY1), averaging 376.7 unique topics (42.1% of MCPEM subcategories). By PGY4, they averaged 565.9 topics (63.2% of MCPEM), representing a 9.9% increase over PGY3. Exposure plateaus generally occurred at 39–41 months, though substantial individual variation was observed, with some residents continuing to acquire new topics until graduation. Annual case volume more than tripled from PGY1 (mean 445.7 encounters) to PGY4 (mean 1,528.4 encounters). Case complexity increased, as evidenced by a decrease in mean ESI score from 2.94 to 2.79 and a rise in high-acuity (ESI 1–2) cases from 16.0% to 30.9%.

Conclusions: NLP analysis of clinical documentation provides a scalable, detailed method for tracking EM resident clinical exposure and progression. Many residents continue to gain new experiences into their fourth year, particularly with higher-acuity cases. These findings suggest that a four-year training model may offer meaningful additional educational value, while also highlighting the importance of individualized assessment given variability in learning trajectories.

(JMIR Preprints 13/08/2025:82326)
DOI: https://doi.org/10.2196/preprints.82326

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Quantifying Emergency Medicine Residency Learning Curves Using Natural Language Processing: A Retrospective Cohort Study

ABSTRACT

Background:

The optimal duration of emergency medicine (EM) residency training remains a subject of national debate, with the Accreditation Council for Graduate Medical Education considering standardizing all programs to four years. However, empirical data on how residents accumulate clinical exposure over time are limited. Traditional measures, such as case logs and diagnostic codes, often fail to capture the breadth and depth of diagnostic reasoning. Natural language processing (NLP) of clinical documentation offers a novel approach to quantify clinical experiences more comprehensively.

Objectives

This study aimed to: (1) quantify how EM residents acquire clinical topic exposure over the course of training; (2) evaluate variation in exposure patterns across residents and classes; and (3) assess changes in workload and case complexity over time to inform the discussion on optimal program length.

Methods

We conducted a retrospective cohort study of EM residents at Stanford Hospital, analyzing 244,255 emergency department encounters from July 1, 2016, to November 30, 2023. The sample included 62 residents across four graduating classes (2020–2023), representing all primary training site encounters where residents served as primary or supervisory providers. Using a retrieval-augmented generation NLP pipeline, we mapped resident clinical documentation to the 895 subcategories of the 2022 Model for Clinical Practice of Emergency Medicine (MCPEM) via intermediate mapping to the SNOMED CT CORE Problem List Subset. We generated cumulative topic exposure curves, quantified the diversity of topic coverage, assessed variability between residents, and analyzed progression in clinical complexity using Emergency Severity Index (ESI) scores and admission rates.

Results

Residents encountered the largest increase in new topics during postgraduate year 1 (PGY1), averaging 376.7 unique topics (42.1% of MCPEM subcategories). By PGY4, they averaged 565.9 topics (63.2% of MCPEM), representing a 9.9% increase over PGY3. Exposure plateaus generally occurred at 39–41 months, though substantial individual variation was observed, with some residents continuing to acquire new topics until graduation. Annual case volume more than tripled from PGY1 (mean 445.7 encounters) to PGY4 (mean 1,528.4 encounters). Case complexity increased, as evidenced by a decrease in mean ESI score from 2.94 to 2.79 and a rise in high-acuity (ESI 1–2) cases from 16.0% to 30.9%.

Conclusions

NLP analysis of clinical documentation provides a scalable, detailed method for tracking EM resident clinical exposure and progression. Many residents continue to gain new experiences into their fourth year, particularly with higher-acuity cases. These findings suggest that a four-year training model may offer meaningful additional educational value, while also highlighting the importance of individualized assessment given variability in learning trajectories.

Keywords: Emergency medicine residency; clinical exposure; learning curves; natural language processing; electronic health records; graduate medical education

INTRODUCTION

The optimal duration of emergency medicine (EM) residency training remains a critical unresolved question in graduate medical education. As the Accreditation Council for Graduate Medical Education (ACGME) considers standardizing all programs to a four-year model, this debate has highlighted a fundamental gap in our understanding: we lack reliable methods to measure how residents accumulate and master clinical experiences across emergency medicine's vast spectrum of presentations.[1] While these clinical encounters form the foundation of physician development and shape future practice patterns, programs have struggled to systematically track and optimize them, even within competency-based educational frameworks.[2–5]

Current methods for measuring clinical exposure in EM suffer from both practical and conceptual limitations. Self-reported case logs, the traditional standard, demonstrate significant error rates due to recall bias.[6,7] Approaches using diagnostic coding systems like International Classification of Diseases, 10th Revision (ICD-10) fundamentally misalign with emergency medicine's paradigm.[8–11] The core work of emergency physicians—evaluating and ruling out life-threatening conditions—often results in non-specific final diagnoses (e.g., "abdominal pain") that mask the complexity of care delivered.[12] Furthermore, we know that evaluating for life-threatening illnesses within the context of abdominal pain can be confounded by mimics like acute coronary syndrome, which are non-specific, have high overlap with other conditions, and while they are considered, may not appear in the final diagnosis but nonetheless expose the resident to evaluating for that condition.

Natural language processing (NLP) of clinical documentation offers a potential breakthrough. By analyzing the comprehensive narrative content of clinical notes—capturing the diagnostic reasoning process—NLP can measure case exposure with greater granularity. When combined with Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)—a clinical terminology system designed to represent complex medical concepts—this approach can systematically document both the breadth of conditions evaluated, and the depth of clinical reasoning employed. This methodological advance aligns with emerging calls for "precision education" in medical training, where analysis of clinical data drives personalized learning optimization.[5,13]

In this study, we used NLP to provide a comprehensive, data-driven analysis of EM resident development. Using electronic health record (EHR) data from a single academic medical center spanning multiple resident cohorts, we pursued three objectives: (1) quantify topic exposure curves and clinical progression, mapping how residents accumulate diagnostic topic exposure over time; (2) examine variation in clinical exposure patterns between individual residents and graduating classes; and (3) analyze the distribution of clinical experiences across presentation types and complexity levels to provide empirical evidence relevant to the debate on optimal training duration.

METHODS

Study Design and Population

We conducted a retrospective analysis of emergency department (ED) encounters at Stanford Hospital between July 1, 2016, and November 30, 2023. Stanford Hospital serves as the primary training site for our residency and is a high-volume academic emergency department and level 1 trauma center. This period captured the primary training site experiences of four resident classes (2020–2023). The study included all EM residents who completed their full four-year training during this period (n=62). Encounters where residents served as either the primary or supervisory resident were included. Our program is a four-year PGY1–4 program where the PGY4 year includes 40 weeks of ED time with a specific emphasis on developing supervisory skills and practicing with graduated responsibility.

Residents also gain clinical experience at two high-volume, high-acuity affiliated sites (Kaiser Santa

Clara and Santa Clara Valley Medical Center) that use a different electronic health record (EHR); this data was not included in our analysis. Depending on the PGY level, these external sites constitute approximately 30–35% of a resident's total ED training time. Encounters from off-service rotations (e.g., intensive care unit, anesthesia, obstetrics) were also excluded. Residents who did not complete the program were excluded from this analysis to ensure the integrity of longitudinal exposure curve construction. The Stanford Institutional Review Board approved this study (IRB 69107), and data were de-identified within a secure institutional data repository to ensure Health Insurance Portability and Accountability Act (HIPAA) compliance before analysis

Data Sources and Variables

We extracted de-identified structured data including patient demographics, Emergency Severity Index (ESI), and disposition status, as well as unstructured data in the form of clinical documentation from Stanford Research Repository (STARR).[14] We focused on note sections capturing resident diagnostic reasoning: history of present illness, medical decision making, and ED course narratives.

Natural Language Processing

We developed a multi-stage NLP pipeline to map the narrative content of resident clinical documentation to the 895 clinical subcategories of the 2022 Model for Clinical Practice of Emergency Medicine (MCPEM).[15] Our approach used a retrieval-augmented generation framework with Google's Gemini 1.5 Flash large language model.[16–18] This involved extracting key clinical concepts from resident notes and mapping them, as an intermediate step, to the Systematized Nomenclature of Medicine Clinical Terms, Clinical Observations, Recordings, and Encoding subset (SNOMED-CT CORE) before final classification into MCPEM topics.[19] This two-stage process was chosen to preserve granular clinical detail while using a standardized clinical terminology.

Validation Methodology

We validated this pipeline through manual review of 500 randomly selected encounters by four board-certified emergency physicians (CP, WD, JH, RK). This process confirmed the high accuracy of our automated approach, with the model's classifications agreeing with the expert consensus 89.7% of the time. The inter-rater reliability among the physician reviewers was substantial (kappa=0.71). A comprehensive description of the NLP architecture, model configuration and validation methodology is provided in **Supplemental Digital Appendix 1.**

Analysis

Our analysis focused on three key aspects of resident development. To appropriately account for the resident as the primary unit of analysis and the clustered nature of the data (i.e., multiple encounters nested within each resident), all encounter-level data were first aggregated to the individual resident level. Statistical comparisons were then performed on this resident-level dataset (N=62).

First, we constructed topic exposure curves by tracking cumulative unique topics over time. Topic exposure rates were calculated using 30-day sliding windows. We defined exposure plateaus as periods where residents encountered fewer than one new topic per 100 patients over three consecutive measurement windows. Second, we examined variation in exposure by analyzing differences in case volumes, topic coverage, and patient acuity. To quantify the equity of exposure distribution among residents within the same PGY level, we calculated Gini coefficients. Originally developed to measure income inequality, the Gini coefficient is a statistic that quantifies the inequality of a frequency distribution, with values ranging from 0 (representing perfect equality) to 1 (representing perfect inequality).[20] In this context, a Gini coefficient of 0 would indicate that all residents in a cohort had the exact same volume of exposure to a given measure (e.g., high-acuity cases), while a value of 1 would indicate that a single resident received all of the exposure and all others received none. This metric provides a standardized way to compare the degree of inter-resident variability across different clinical domains and training years. Third, we tracked clinical complexity progression using ESI scores and admission rates as proxies.

Analyses were performed using R version 4.3.1 (R Foundation for Statistical Computing) and Python version 3.11. Statistical comparisons between classes used Kruskal-Wallis tests with eta-squared effect sizes. A *P* value of <.05 was considered statistically significant.

RESULTS

Resident Cohort

Our analysis included the primary-site training experiences of 62 EM residents from four graduating classes (2020–2023). Over the course of their training, this cohort managed 244,255 patient encounters, representing 133,748 unique patients Detailed demographic and clinical characteristics of the patient encounters are provided in Table 1.

Topic Exposure Progression and Inter-resident Variation

EM residents demonstrated a clear, progressive acquisition of clinical topic exposure throughout training (Figures 1–4). The most rapid new topic exposure occurred during postgraduate year 1 (PGY1), with residents encountering a mean of 376.7 unique topics (42.0% of the MCPEM subcategories). However, PGY1 was also the year of greatest inter-resident variability in topic coverage (coefficient of variation, CV=9.1%). As residents progressed through training, the variation in total topic exposure decreased (PGY4 CV=2.8%).

Exposure coverage continued to increase in subsequent years, reaching a mean of 447.6 topics in PGY 2 (50.0% of MCPEM), 515.0 topics in PGY3 (57.5% of MCPEM), and 565.9 topics in PGY4 (63.2% of MCPEM). Exposure plateaus, defined as periods with minimal new topic exposure, typically occurred in the fourth year of training, for example The Class of 2023 reached plateaus at a mean of 39.8 (SD 3.0) months (mean 3,268 encounters). However, individual variation was substantial: some residents plateaued as early as 38.2 months (2,742 encounters), while others continued to encounter new topics through their final month, with final topic coverage ranging from 59.4% to 67.3% of all possible topics among PGY4 residents.

Progression of Clinical Workload and Complexity

In parallel with increasing topic exposure, residents demonstrated a progressive increase in their clinical workload and the complexity of cases managed (Table 2). Annual case volumes more than tripled from PGY1 (mean 445.7 encounters) to PGY4 (mean 1,528.4 encounters). This growth in volume was accompanied by increasing case acuity, as mean ESI scores progressively decreased from 2.94 in PGY1 to 2.79 in PGY4. Correspondingly, the proportion of high-acuity patients (ESI 1-2) managed by residents increased from 16.0% in PGY1 to 30.9% in PGY4, and admission rates rose from 31.0% to 37.5% over the same period. Of note, the CV for both clinical volume and admission rates followed a U-shaped pattern, decreasing from PGY1 to PGY3 before increasing again in PGY4.

Distribution of Clinical Experiences

Residents' exposure to the 895 distinct clinical topics followed a consistent pattern: they encountered a small core of presentations (5.5%) more than 100 times each, a larger set (31.7%) between 10 and 100 times, and the majority (62.9%) fewer than 10 times. Topic distribution showed moderate inequality (mean Gini coefficient=0.611), with consistency between graduating classes *(P*=.61). High-acuity case exposure was more unequally distributed (Gini=0.292) than overall case volumes (Gini=0.117).

DISCUSSION

Our analysis of 62 emergency medicine residents across four years of training revealed distinct and progressive patterns in the arc of their clinical experience. By using a novel NLP

methodology on over 244,000 clinical encounters, we found that residents demonstrate a rapid acquisition of topic exposure in their first year, which continues, albeit at a slower rate, deep into their fourth year. Importantly, this continued exposure occurs in the context of increasing clinical complexity and based on our program's structure, escalating supervisory responsibility. These findings provide empirical evidence that can inform the national debate on the optimal length of EM training and highlight the potential for data-driven, precision education.

The observed pattern of topic exposure—rapid initial acquisition followed by a plateau—aligns with the power-law "experience curves" documented in medical education by Pusic et al.[21] However, it is critical to note that case exposures are merely the substrate for learning, and are not all equal in value towards developing competence. The conversion of these experiences into durable competence is a complex process mediated by the *deliberate* components of Ericsson's theory of deliberate practice, such as feedback, reflection, and coaching.[22–24] Indeed, recent research has shown that even established competency measures like milestones may not directly correlate with early-career patient outcomes, cautioning against a simple equation of more exposure with more competence.[25] Therefore, the primary value of the exposure data we present lies in its ability to serve as a powerful objective input for these established educational frameworks. At our institution, for example, these components are formalized through a resident coaching program and a quarterly Clinical Competency Committee (CCC), systems for which this objective exposure data can provide a more precise foundation for assessment and goal-setting.[2,26]

A central question for the EM as a specialty is whether a three- or four-year training model is optimal. Our data provide empirical evidence relevant to this debate. The finding that residents

continue to acquire a mean of 50.9 new core topics in their PGY4 year, representing a 9.9% increase over PGY3, suggests that the fourth year offers more than just redundant experience. This quantitative increase is accompanied by a significant qualitative shift: PGY4 residents manage a higher proportion of high-acuity patients (ESI 1–2) and cases requiring hospitalization. This exposure to a more complex and challenging case-mix, as supported by work from Lam et al. and Liu et al., is critical for developing the advanced diagnostic reasoning necessary for independent practice.[10,27]

Our analysis also revealed significant inter-resident variation, particularly in the timing of exposure plateaus and in the experience with high-acuity cases. This aligns with findings from other specialties and supports a more personalized view of competency development.[28] This variability is likely driven by a combination of systemic factors and resident choice. We propose framing this variability within the concept of "warranted versus unwarranted variation" as described by Holmboe and Kogan.[29] While some variation is an expected and even desirable feature of individualized learning, our NLP-based tool provides a mechanism for programs to identify potentially unwarranted gaps in exposure to core experiences. Notably, the decreasing coefficient of variation in total topic coverage from PGY1 (9.1%) to PGY4 (2.8%) suggests that a longer training duration may lead to a more standardized and equitable clinical experience among graduates.

The U-shaped pattern observed in the variability of clinical volume and admission rates is an intriguing finding. We hypothesize this reflects the evolving roles within our program: PGY1 residents have variable schedules due to numerous off-service rotations; PGY2 and PGY3 residents assume more structured supervisory roles within a pod system, which may standardize their workflow and decrease variability; finally, in PGY4, residents are granted

greater autonomy to run a pod independently while also supervising junior residents, potentially allowing individual practice styles to reemerge at increasing variability.

A key finding of this study is that even after four years, residents were not exposed to over a third of the topics listed in the MCPEM. This highlights the fundamental tension between the prescribed curriculum and the reality of clinical practice. Achieving 100% topic coverage is likely an unachievable goal for any single residency program. The true aim of training is not encyclopedic exposure but rather developing the core competencies and adaptive expertise required for safe, independent practice and effective lifelong learning. Our NLP-based methodology offers a powerful, dual-pronged approach to addressing this gap. At the local program level, it serves as a diagnostic tool, enabling educators to identify and amend exposure gaps through targeted interventions like simulation. More broadly, the scalability of this approach presents an opportunity to transform the specialty's understanding of its own work. If applied across multiple institutions, this method could create a dynamic, data-driven map of the actual clinical practice of EM, providing an evidence base for future revisions of the MCPEM. This would allow the model to better reflect the true prevalence and complexity of conditions encountered in contemporary practice, ensuring more authentic alignment between what is taught, what is tested, and what is practiced. Nonetheless, there will likely always remain a set of high-acuity, low-frequency conditions for which training programs must prescribe exposure through didactics, as real-world encounters will be too rare to ensure universal competence.

A key advantage of this methodology lies in its practical implementation. Traditional NLP pipelines often require a substantial upfront investment in manual data annotation for model training and specialized expertise. In contrast, our retrieval-augmented generation approach takes advantage of a pre-trained large language model that required no model-specific

training, dramatically reducing implementation complexity. To offer a concrete sense of financial feasibility, the total cost for the application programming interface calls to process all 244,255 clinical encounters for this study was approximately $180, demonstrating the financial accessibility of this approach for programs seeking to adopt data-driven educational strategies.

This study has several limitations. As a single-institution study, our findings may not generalize to three-year programs, particularly three-year programs or those in different clinical settings. A primary limitation is that our analysis excludes data from additional training sites, which constitute a significant proportion of resident training (approximately 30–35% of ED time, depending on the PGY level), and from off-service rotation where key procedural and critical care exposures occur (e.g., anesthesia, intensive care unit). The training period for our cohorts also overlapped with the COVID-19 pandemic, which may have influenced case volumes and mix, though the remarkable consistency of exposure patterns we observed across the four classes mitigates this concern. Finally, while our NLP approach achieved high accuracy in classifying clinical encounters (89.7% agreement with physician review), this methodology relies on the comprehensiveness of resident documentation, which may vary between individuals and over time.

In conclusion, our analysis reveals both consistent patterns in resident clinical exposure and substantial individual variation in topic exposure trajectories. The finding that many residents continue to encounter new clinical topics into their fourth year provides empirical evidence for the potential educational value of a four-year training model. Natural language processing of clinical documentation offers emergency medicine programs a powerful and accessible tool to objectively measure and optimize resident clinical experiences based on actual exposure patterns, moving beyond traditional metrics to foster a more precise, data-informed, and equitable approach to graduate medical education.

Acknowledgements: None.

Funding: None.

Conflicts of Interest: None.

Data Availability: The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Authors' Contributions: Conceptualization: CP, CR. Methodology: CP, JH, RK, CR. Software and Data Curation: CP. Formal Analysis: CP. Writing – Original Draft: CP, JH, RK, CR, WD. Writing – Review & Editing: CP, JH, RK, WD, CR.

Abbreviations: NLP: natural language processing; EM: emergency medicine; MCPEM: Model for Clinical Practice of Emergency Medicine; SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms; ED: Emergency Department; PGY: Postgraduate year; ESI: Emergency Severity Index; CV: coefficient of variation

REFERENCES

1. ACGME Program Requirements for Graduate Medical Education in Emergency Medicine. 2025. Available from: https://www.acgme.org/globalassets/pfassets/reviewandcomment/2025/110_emergencymedicine_rc_02122025.pdf [accessed Mar 17, 2025]

2. Sebok-Syer SS, Shepherd L, McConnell A, Dukelow AM, Sedran R, Lingard L. “EMERGing” Electronic Health Record Data Metrics: Insights and Implications for Assessing Residents’ Clinical Performance in Emergency Medicine. Chan T, editor. AEM Education and Training 2021 Apr;5(2):e10501. doi: 10.1002/aet2.10501

3. Sirovich BE, Lipner RS, Johnston M, Holmboe ES. The association between residency training and internists’ ability to practice conservatively. JAMA Intern Med 2014 Oct;174(10):1640–1648. PMID:25179515

4. Chen C, Petterson S, Phillips R, Bazemore A, Mullan F. Spending Patterns in Region of Residency Training and Subsequent Expenditures for Care Provided by Practicing Physicians for Medicare Beneficiaries. JAMA 2014 Dec 10;312(22):2385–2393. doi: 10.1001/jama.2014.15973

5. Burk-Rafel J, Drake CB, Sartori DJ. Characterizing Residents’ Clinical Experiences—A Step Toward Precision Education. JAMA Netw Open 2024 Dec 18;7(12):e2450774. doi: 10.1001/jamanetworkopen.2024.50774

6. Langdorf MI, Strange G, Macneil P. Computerized tracking of emergency medicine resident clinical experience. Annals of Emergency Medicine 1990 July;19(7):764–773. doi: 10.1016/S0196-0644(05)81700-7

7. Nagler J, Harper MB, Bachur RG. An Automated Electronic Case Log: Using Electronic Information Systems to Assess Training in Emergency Medicine. Academic Emergency Medicine 2006 July;13(7):733–739. doi: 10.1197/j.aem.2006.02.010

8. Rhee DW, Reinstein I, Jrada M, Pendse J, Cocks P, Stern DT, Sartori DJ. Mapping hospital data to characterize residents’ educational experiences. BMC Med Educ 2022 Dec;22(1):496. doi: 10.1186/s12909-022-03561-x

9. Rhee DW, Chun JW, Stern DT, Sartori DJ. Experience and Education in Residency Training: Capturing the Resident Experience by Mapping Clinical Data. Academic Medicine 2022 Feb;97(2):228–232. doi: 10.1097/ACM.0000000000004162

10. Lam ACL, Tang B, Liu C, Ismail MF, Roberts SB, Wankiewicz M, Lalwani A, Schumacher D, Kinnear B, Verma AA, Razak F, Wong BM, Ginsburg S. Variation in Case Exposure During Internal Medicine Residency. JAMA Netw Open 2024 Dec 18;7(12):e2450768. doi: 10.1001/jamanetworkopen.2024.50768

11. Lam AC, Tang B, Lalwani A, Verma AA, Wong BM, Razak F, Ginsburg S. Methodology paper for the General Medicine Inpatient Initiative Medical Education Database (GEMINI MedED): a retrospective cohort study of internal medicine resident case-mix, clinical care and patient outcomes. BMJ Open 2022 Sept;12(9):e062264. doi: 10.1136/bmjopen-2022-062264

12. Bischof JJ, Emerson G, Mitzman J, Khandelwal S, Way DP, Southerland LT. Does the Emergency Medicine In-training Examination Accurately Reflect Residents’ Clinical

Experiences? Wolff M, editor. AEM Education and Training 2019 Oct;3(4):317–322. doi: 10.1002/aet2.10381

13. Triola MM, Burk-Rafel J. Precision Medical Education. Academic Medicine 2023 July;98(7):775–781. doi: 10.1097/ACM.0000000000005227

14. Weber SC, Pallas J, Olson G, Love D, Malunjkar S, Boosi S, Loh E, Datta S, Ferris TA. Compliant Self Service Access to Secondary Use Clinical Data at Stanford Medicine. arXiv; 2024. doi: 10.48550/arXiv.2412.04248

15. Beeson MS, Bhat R, Broder JS, Gaeta TJ, Janssen A, Karl ER, Lo BM, Moll J, Oh L, Patel V, Touma L, Barton MA, Keehbauch JN, 2022 EM Model Review Task Force; Melissa A. Barton, M.D.(∗), and Julia N. Keehbauch(∗), American Board of Emergency Medicine. The 2022 Model of the Clinical Practice of Emergency Medicine. J Emerg Med 2023 June;64(6):659–695. PMID:37244783

16. Gemini models | Gemini API | Google AI for Developers. Available from: https://ai.google.dev/gemini-api/docs/models [accessed Mar 17, 2025]

17. Klang E, Tessler I, Apakama DU, Abbott E, Glicksberg BS, Arnold M, Moses A, Sakhuja A, Soroush A, Charney AW, Reich DL, McGreevy J, Gavin N, Carr B, Freeman R, Nadkarni GN. Assessing Retrieval-Augmented Large Language Model Performance in Emergency Department ICD-10-CM Coding Compared to Human Coders. medRxiv 2024 Oct 17;2024.10.15.24315526. PMID:39484238

18. D’Oosterlinck K, Khattab O, Remy F, Demeester T, Develder C, Potts C. In-Context Learning for Extreme Multi-Label Classification. arXiv; 2024. doi: 10.48550/arXiv.2401.12178

19. The CORE Problem List Subset of SNOMED CT®. U.S. National Library of Medicine; Available from: https://www.nlm.nih.gov/research/umls/Snomed/core_subset.html [accessed Mar 17, 2025]

20. Thomas V, Wang Y, Fan X. Measuring Education Inequality: Gini Coefficients of Education. Rochester, NY: Social Science Research Network; 2001. Available from: https://papers.ssrn.com/abstract=258182 [accessed June 12, 2025]

21. Pusic MV, Kessler D, Szyld D, Kalet A, Pecaric M, Boutis K. Experience Curves as an Organizing Framework for Deliberate Practice in Emergency Medicine Learning. Mycyk MB, editor. Academic Emergency Medicine 2012 Dec;19(12):1476–1480. doi: 10.1111/acem.12043

22. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004 Oct;79(10 Suppl):S70-81. PMID:15383395

23. Natesan S, Jordan J, Sheng A, Carmelli G, Barbas B, King A, Gore K, Estes M, Gottlieb M. Feedback in Medical Education: An Evidence-based Guide to Best Practices from the Council of Residency Directors in Emergency Medicine. West J Emerg Med 2023 May 5;24(3):479–494. PMID:37278777

24. Fredette J, Michalec B, Billet A, Auerbach H, Dixon J, Poole C, Bounds R. A qualitative assessment of emergency medicine residents’ receptivity to feedback. AEM Educ Train 2021 Aug 1;5(4):e10658. PMID:34527849

25. Kendrick DE, Thelen AE, Chen X, Gupta T, Yamazaki K, Krumm AE, Bandeh-Ahmadi H, Clark

M, Luckoscki J, Fan Z, Wnuk GM, Ryan AM, Mukherjee B, Hamstra SJ, Dimick JB, Holmboe ES, George BC. Association of Surgical Resident Competency Ratings With Patient Outcomes. Acad Med 2023 July 1;98(7):813–820. PMID:36724304

26. Sebok-Syer SS, Shaw JM, Sedran R, Shepherd L, McConnell A, Dukelow AM, Syer MD, Lingard L. Facilitating Residents' Understanding of Electronic Health Record Report Card Data Using Faculty Feedback and Coaching. Academic Medicine 2022 Nov;97(11S):S22–S28. doi: 10.1097/ACM.0000000000004900

27. Zhou L, Cai C, Wu R, Qi Y. Effectiveness of scaffolded case-based learning in anesthesiology residency training: a randomized controlled trial. BMC Med Educ 2025 May 8;25(1):672. PMID:40340724

28. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, Harris P, Glasgow NJ, Campbell C, Dath D, Harden RM, Iobst W, Long DM, Mungroo R, Richardson DL, Sherbino J, Silver I, Taber S, Talbot M, Harris KA. Competency-based medical education: theory to practice. Med Teach 2010;32(8):638–645. PMID:20662574

29. Holmboe ES, Kogan JR. Will Any Road Get You There? Examining Warranted and Unwarranted Variation in Medical Education. Acad Med 2022 Aug 1;97(8):1128–1136. PMID:35294414

TABLES

Table 1: Demographic and Clinical Characteristics of Emergency Department Patient Encounters Managed by Emergency Medicine Residents at Stanford Hospital, 2016-2023

Characteristic	Value
Patient Demographics	
Total unique patients	133,748
Mean age, years (SD)	47.2 (25.7)
Gender	
Female	69,015 (51.6%)
Male	64,666 (48.4%)
Unknown	67 (0.0%)
Race	
White	55,948 (41.8%)
Other/Multiple	43,939 (32.9%)
Asian	22,994 (17.2%)
Black	6,418 (4.8%)
Pacific Islander	2,469 (1.9%)
Unknown	1,564 (1.17%)
Native American	414 (0.3%)
Ethnicity	
Non-Hispanic	96,965 (72.5%)
Hispanic/Latino	35,132 (26.3%)
Unknown	1,649 (1.2%)
Insurance	
Commercial	48,894 (36.6%)
Medicare	30,357 (22.7%)
Medicaid/Medi-Cal	33,632 (25.2%)
Unknown	11,617 (8.7%)
Other	9,242 (6.9%)
Primary Language	
English	110,165 (82.4%)
Spanish	16,371 (12.2%)
Chinese Languages	2,257 (1.7%)
Southeast Asian	1,513 (1.1%)
Other	3,339 (2.5%)
Interpreter Services	
Interpreter Needed	20,483 (15.3%)
No Interpreter Needed	113,162 (84.6%)
Unknown	103 (0.1%)
Encounter Characteristics	
Total patient encounters	244,255
Encounters per resident, mean [95% CI]	3,940 [3,794–4,086]
Admission rate, %	35.7%
Length of stay, hours (median, IQR)	4.8 [3.1–7.2]

Emergency Severity Index	
1	2,759 (1.2%)
2	57,870 (24.1%)
3	157,556 (65.7%)
4	19,766 (8.2%)
5	1,894 (0.8%)
Missing	4,410 (1.8%)

Abbreviations: SD, standard deviation; CI, confidence interval; IQR, interquartile range

Table 2: Progression of Clinical Experience by Postgraduate Year (PGY) for Emergency Medicine Residents (n=62), Stanford Hospital, 2016-2023

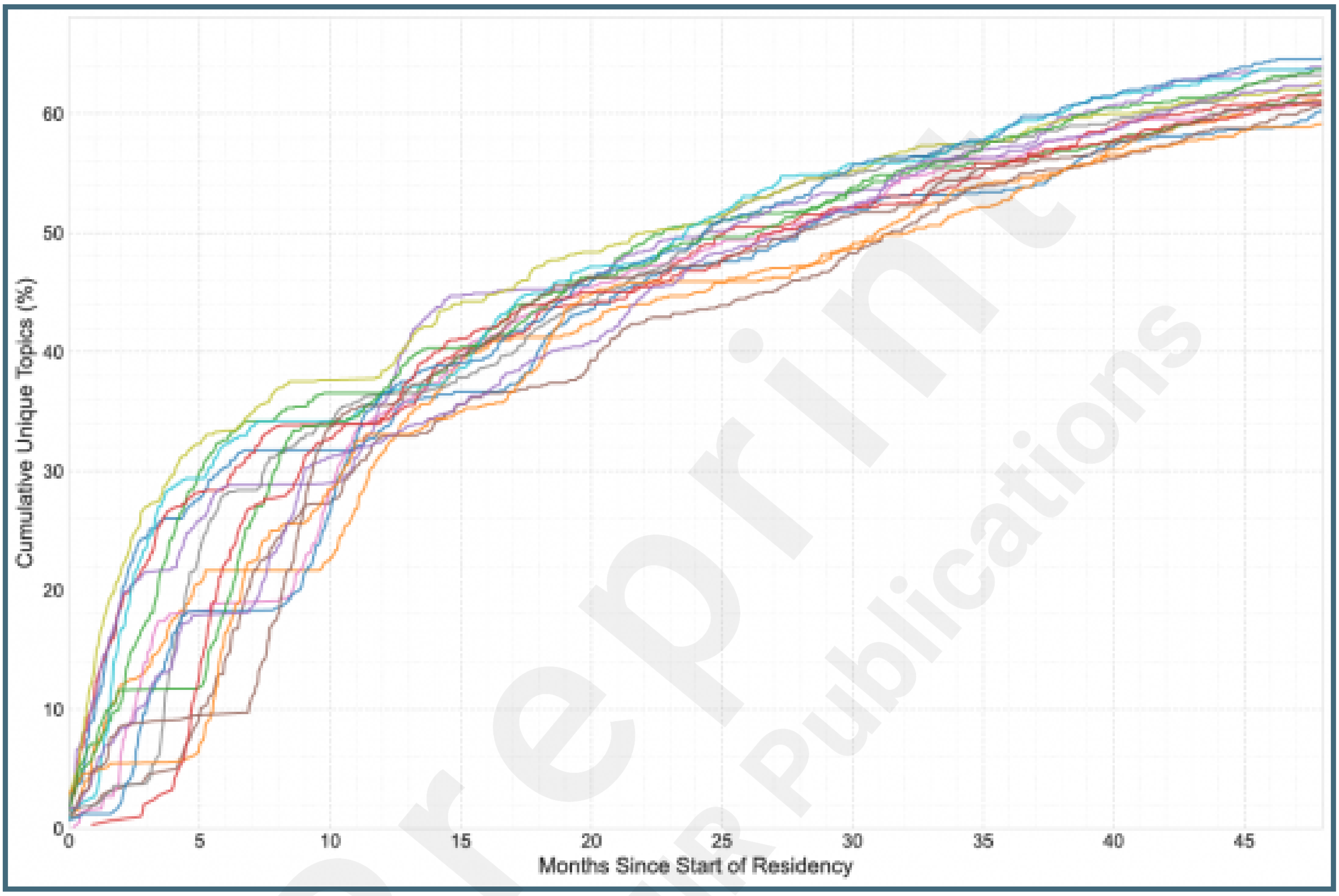
	PGY1	PGY2	PGY3	PGY4
Annual Clinical Volume and Complexity				
Number of encounters, mean [95% CI]	445.7 [417.6–473.7]	772.1 [738.5–805.8]	1193.4 [1138.5–1248.2]	1528.4 [1429.1–1627.8]
ESI score, mean [95% CI]	2.94 [2.93–2.96]	2.87 [2.85–2.88]	2.85 [2.84–2.87]	2.79 [2.76–2.81]
High-acuity cases (ESI 1–2), %	16.0%	20.3%	24.5%	30.9%
Admission rate, % [95% CI]	31.0% [29.7–32.4]	38.3% [37.3–39.4]	35.5% [34.6–36.3]	37.5% [36.3–38.8]
Cumulative Topic Exposure				
Cumulative unique MCPEM topics covered, mean [95% CI]	376.7 [367.9–385.5]	447.6 [442.2–453.0]	515.0 [510.9–519.0	565.9 [561.9–569.9]
Cumulative total % of MCPEM covered	41.1%	50.0%	57.5%	63.2%
New topics encountered each year, mean	376.7	71.0	67.3	50.9
Inter-resident variability (CV)				
CV for clinical volume	24.6%	17.0%	18.0%	25.4%
CV for admission rate	16.4%	11.1%	9.5%	13.4%
CV for topic coverage	9.1%	4.7%	3.1%	2.8%

CI, confidence interval; CV, coefficient of variation; ESI, Emergency Severity Index; MCPEM, Model for Clinical Practice of Emergency Medicine

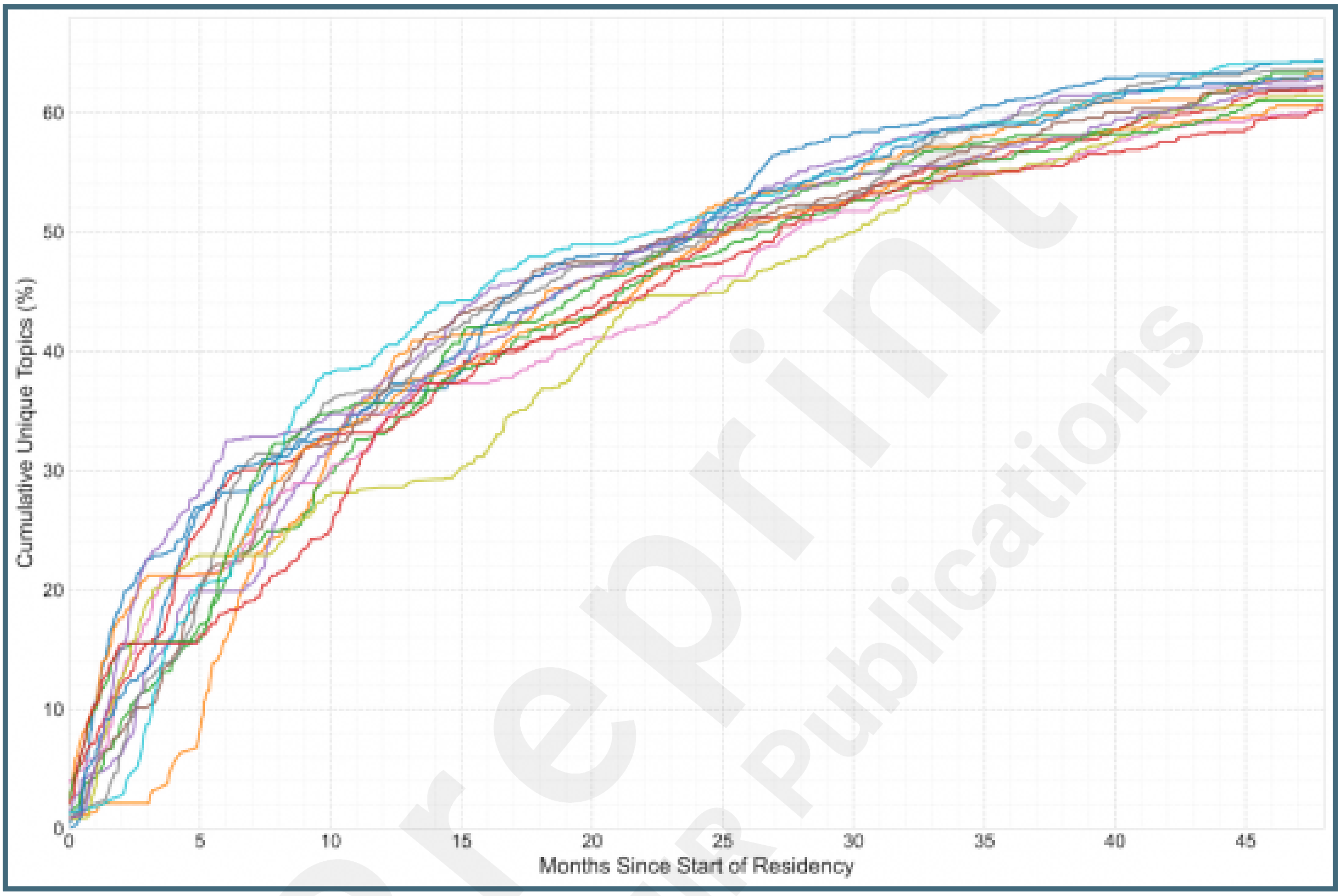
Supplementary Files

Figures

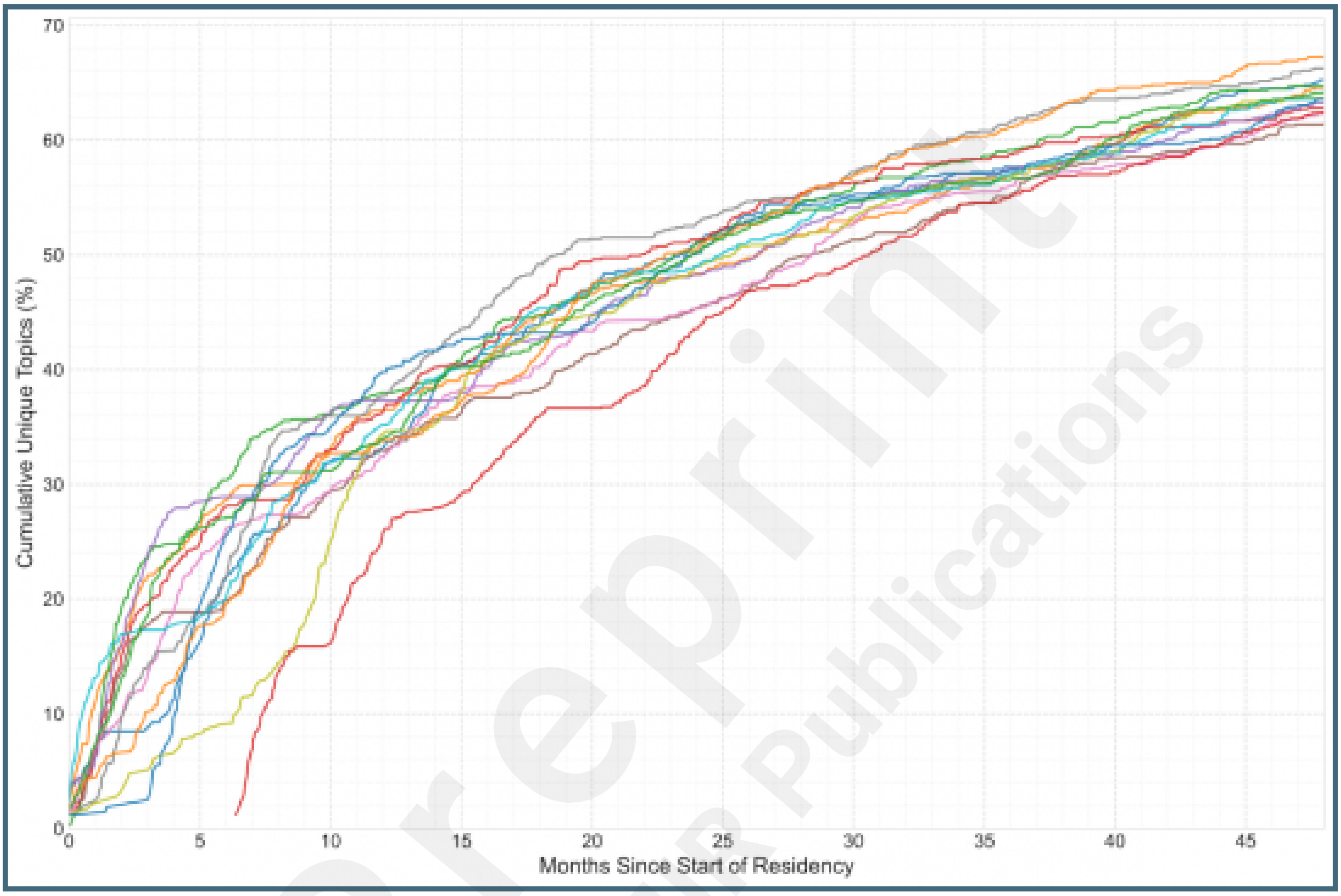
Cumulative clinical topic exposure curve for the Emergency Medicine graduating class of 2020 at Stanford Hospital, 2016–2020 (n=15). The x-axis represents months of training, and the y-axis represents the mean cumulative number of unique clinical topics from the Model for Clinical Practice of Emergency Medicine (MCPEM) encountered by residents. The curve does not reach the total of 895 MCPEM topics, indicating that no resident achieved 100% topic exposure during their training at the primary academic site.



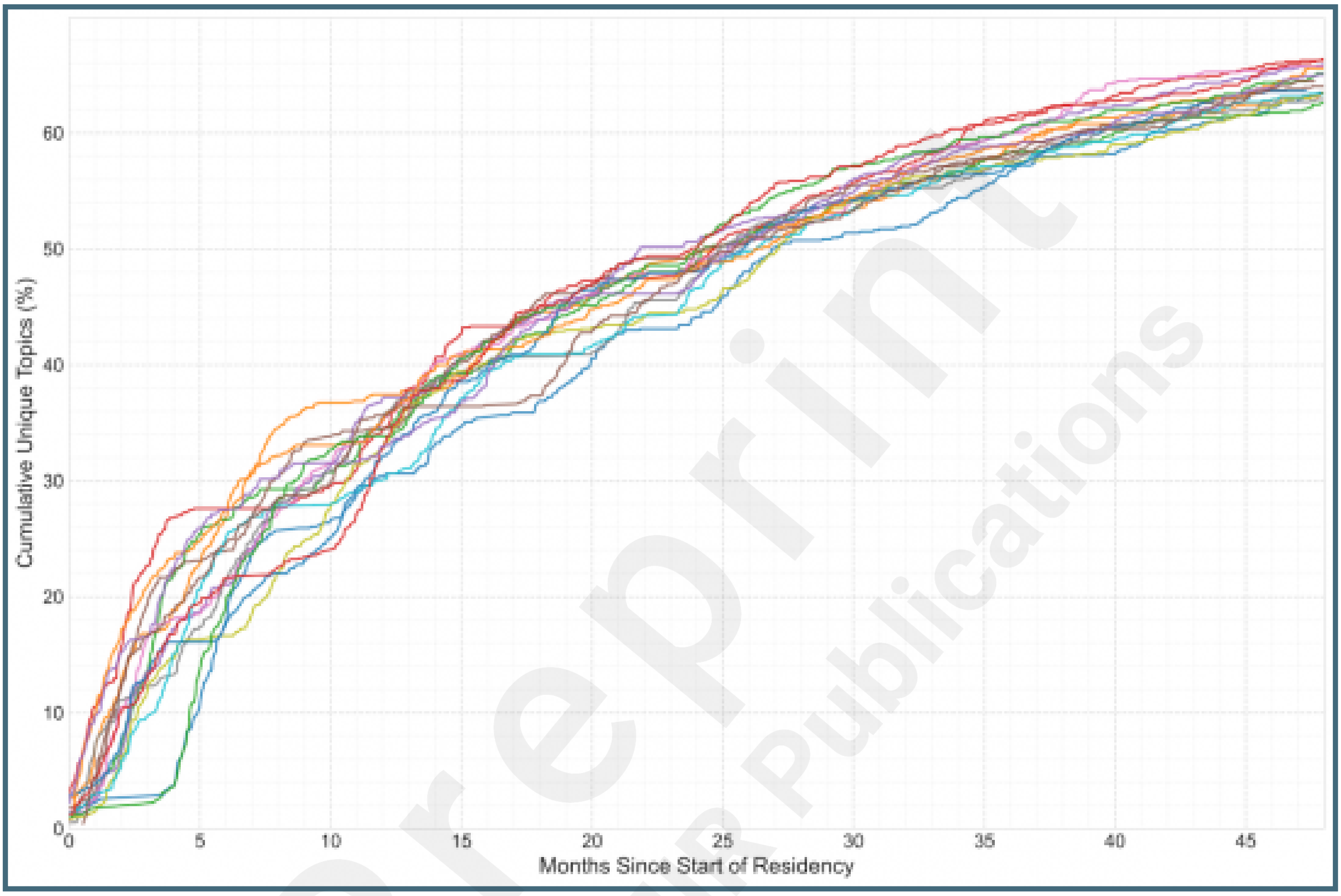
Cumulative clinical topic exposure curve for the Emergency Medicine graduating class of 2021 at Stanford Hospital, 2017–2021 (n=16). The x-axis represents months of training, and the y-axis represents the mean cumulative number of unique clinical topics from the Model for Clinical Practice of Emergency Medicine (MCPEM) encountered by residents. The curve does not reach the total of 895 MCPEM topics, indicating that no resident achieved 100% topic exposure during their training at the primary academic site.



Cumulative clinical topic exposure curve for the Emergency Medicine graduating class of 2022 at Stanford Hospital, 2018–2022 (n=15). The x-axis represents months of training, and the y-axis represents the mean cumulative number of unique clinical topics from the Model for Clinical Practice of Emergency Medicine (MCPEM) encountered by residents. The curve does not reach the total of 895 MCPEM topics, indicating that no resident achieved 100% topic exposure during their training at the primary academic site.



Cumulative clinical topic exposure curve for the Emergency Medicine graduating class of 2023 at Stanford Hospital, 2019–2023 (n=16). The x-axis represents months of training, and the y-axis represents the mean cumulative number of unique clinical topics from the Model for Clinical Practice of Emergency Medicine (MCPEM) encountered by residents. The curve does not reach the total of 895 MCPEM topics, indicating that no resident achieved 100% topic exposure during their training at the primary academic site.



Multimedia Appendixes

Supplementary digital appendix.
URL: http://asset.jmir.pub/assets/4f2f4ce666839a67cf3265dbea2a46e4.docx