Dataset for Hospital Length of Stay Prediction as a model to foster Artificial Intelligence innovation

Alejandro Hernández-Arango MD Msc [1,3] , Jesús Francisco Vargas-Bonilla[4], Hernan F Garcia.[4] Valery Quiroga [2]

1. Chief Medical information Officer, Hospital Alma Máter; Medellin Colombia
2. Professor of Engineer Faculty University of Antioquia
3. Professor of Medicine Faculty, Internal Medicine Division, University of Antioquia
4. SISTEMIC Research Group, University of Antioquia

Link to the dataset:

- Data available at - https://bit.ly/464ugPg

1. Introduction

One of the main applications of artificial intelligence (AI) in medicine is to predict the length of stay (LOS) of patients in hospitals, which can help optimize resource allocation and improve patient outcomes. However, predicting LOS accurately and reliably requires a large and diverse dataset that reflects the complexity and variability of the medical context. In this paper, we present “GenStay” dataset from a massive collection of curated discharge data based electronic health records (EHR) (GHIPS) in Hospital Alma Máter, Colombia. We follow the principles of data fairness(Wilkinson et al., 2019) and standing together (*STANDING together - 1 - documentation recommendations*, n.d.), two initiatives that aim to promote ethical and inclusive practices in AI, into the curation and evaluation of the dataset.

Keywords: Lenght of stay, Dathaton, Dataset, Artificial intelligence, Middle-income country, prognosis

Methods

The GenStay dataset contains electronic health records (EHR) data from January 2022 to August 2023 of patients who were discharged from Alma Mater Hospital, a tertiary care hospital in Colombia. The dataset has 67 variables, including demographic, clinical, and administrative information. The main variable of interest is DX_PRINCIPAL, which is the primary diagnosis code according to the ICD-10 Colombian version (DX means ICD-10). The dataset also provides the description (D_DX_PPAL), the chapter (D_Capitulo_CIE10), the category (Categoria), and the subchapter (D_subcapitulo_CIE10) of the primary diagnosis. Other variables related to the diagnosis are DX_MEDICO, DX_R1, DX_R2, DX_R3, DX_F3, DX_F3_R1, DX_F3_R2, DX_F3_R3, DX_pre_cx, DX_pos_cx, and DX_MUERTE, which are the medical diagnosis, the first to third related diagnoses, the first to third final diagnoses, the pre-surgery diagnosis, the post-surgery diagnosis, and the death diagnosis, respectively. Each of these variables has a corresponding description, subchapter, and chronicity classification. The dataset also includes variables such as gender (vGenero), age at discharge (EdadEgreso), year of discharge (Año_Salida), month of discharge (MES), type of affiliation (Tipo_Afiliacion), insurance company (Aseguradora), contract modality (Modalidad_Contrato), hospitalization unit (Hospitalizacion), block (bloque), floor (piso), type of admission (Tipo_Internacion), life cycle stage (Ciclo_Vital), specialty group (ESPECIALIDAD_GRD), admission service (SERVICIO_ADMITE), discharge service (MODALIDAD), discharge type (Tipo_egreso), hours of discharge (HorasdeAlta), days of stay in service (DiasEstanciaServicio), days of stay in hospital (DiasEstanciaClinica), exact days of stay in hospital (DiasEstanciaClinica_exacto), physician ID (ID_Medico_registra), diagnosis type (NombreTipoDiagnostico), transfusions (Transfusiones), antibiotics (Antioboticos), and intensive care unit admission (UCI_UCE). The GenStay dataset is intended for researchers and practitioners who are interested in exploring the factors that influence the length of stay and the outcomes of patients with different diagnoses using or combinating generative artificial intelligence methods. Table 1 Provide a summary of the most important clinical variables.

The GenStay dataset contains data from Hospital Alma Mater, which is available under a confidentiality agreement. The dataset was released on November 2, 2021, and has the version number 1.0. The licensing arrangements are CC BY-NC-ND. The data custodian is the Hospital Alma Mater and Engineer Faculty of University of Antioquia, which can be contacted for any inquiries or issues regarding the dataset. The dataset follows the FAIR principles, which are Findability, Accessibility, Interoperability, and Reusability. The curation team has taken some steps to address these issues, such as consulting with experts and stakeholders from different backgrounds and perspectives, such as medical professionals and engineers, to ensure that the dataset is relevant, useful, and ethical for the intended purposes and be confident about applying rigorous and consistent standards and methods for data extraction, processing, and quality control, to ensure that the dataset is accurate, complete, and reliable so the data extraction was done with the approval and confidentiality agreement of the hospital, which may limit the access and use of the data by other parties with other purposes. The main goal of the data curation was to de-identify the data to protect the privacy of the patients and the hospital staff.

Discussion

The dataset is intended for researchers and practitioners who are interested in exploring the factors that influence the length of stay (Stone et al., 2022)((Xu et al., 2022) (Chrusciel et al., 2021) and the outcomes of patients with different diagnoses using or combination generative artificial intelligence methods.

In terms of innovations in world health care delivery. there are some challenges and opportunities for transforming health care systems in the next decade, driven by factors such as digital health, consumerism, financial constraints(Zimlichman Eyal, Nicklin Wendy, Aggarwal Rajesh, Bates David W., 2021) The future health systems will be more person-centered, prevention-oriented, community-based, and learning-oriented . The role and implications of new payment models, provider roles, and nontraditional competitors and partners will be a key changing factor Healthcare inequality in Colombia its a big problem; wealth, urban residence and type of health insurance, are the most significant predictors and contributors of healthcare utilization and pro-rich inequality, especially for preventive and outpatient care(Garcia-Ramirez et al., 2020)

One way to address this important perspective can be highlighting the importance of interdisciplinary research and innovation in AI , which can draw from and influence many other fields of science and inspire new AI methods and algorithms(Kusters et al., 2020). Therefore, implement a successfully innovation framework in a middle income country could be challenging , but we believe is not impossible.

In this *structured literature review by Secinaro et al* (Secinaro et al., 2021) *of* 288 peer-reviewed papers with bibliometric analysis to explore artificial intelligence in healthcare from various perspectives. The authors reveals five research clusters: health services management, predictive medicine, patient data, diagnostics, and clinical decision-making. The USA tops the list of countries with the maximum number of articles on the topic (215). It is followed by China (83), the UK (54), India (51), Australia (54), and Canada (32). The also Highlight the role of international research collaborations. Colombia doesn’t appear in the list.

AI can help with the challenges of Colombian health care system, this issues could be addressed conducting frameworks between engineers and physicians academics with interdisciplinary point of view that calls for more ethical, transparent and accountable AI practices given the importance of explainability, bias awareness, evaluation methods and regulation for AI development and decision making in healthcare.(Murphy et al., 2021)

One of the limitations of GenStay dataset is that it does not include information about gender identity, race, ethnicity, socioeconomic status, and sexual orientation of the participants. These attributes are relevant for health research, as they are associated with different health outcomes and interactions with wider social factors.(Veenstra, 2011)The lack of information on these attributes can introduce bias in this dataset and limit the generalizability in this typer of population of the findings derived from this data.(Ruprecht et al., 2021)

Other limitation is the curation process could introduce some information loss or distortion in the dataset because some details or features may have been removed or modified to prevent re-identification.

Other identified bias in the data is that the data comes from a single hospital source, which may not represent the general population of patients with similar conditions or needs. This could introduce selection bias in the dataset, because the hospital has some specific characteristics or policies that differ from other hospitals.(Nguyen et al., 2021)

FInally the GenStay dataset contains information of vulnerable population groups, such as patients with severe disabilities and geriatric patients. These groups may have specific health needs and challenges that are not adequately addressed by the current dataset. We recommend that future studies include data on these groups and explore their health outcomes and experiences in relation to the wider social context.(Buttery et al., 2022)

6. Conclusion

In this paper, we introduce the GenStay dataset, a rich and diverse collection of electronic health records (EHR) from Hospital Alma Máter in Colombia. The dataset covers a period of 20 months, from January 2022 to August 2023, and contains various types of information, such as demographic, clinical, and administrative data and ICD-10 diagnosis codes. We curated the dataset with a special attention to data fairness and inclusivity, following ethical AI guidelines.

The GenStay dataset offers a unique opportunity for applying AI to predict patient length of stay (LOS) in hospitals. LOS prediction is a crucial task for hospital management, as it can help optimize resource utilization and improve patient satisfaction. The GenStay dataset captures the complexity and diversity of the real-world medical scenario, which can enhance the validity and robustness of AI predictions.

Bibliography

Buttery, S. C., Philip, K. E. J., Alghamdi, S. M., Williams, P. J., Quint, J. K., & Hopkinson, N. S. (2022). Reporting of data on participant ethnicity and socioeconomic status in high-impact medical journals: a targeted literature review. *BMJ Open*, *12*(8), e064276.

Chrusciel, J., Girardon, F., Roquette, L., Laplanche, D., Duclos, A., & Sanchez, S. (2021). The prediction of hospital length of stay using unstructured data. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01722-4

Garcia-Ramirez, J., Nikoloski, Z., & Mossialos, E. (2020). Inequality in healthcare use among older people in Colombia. *International Journal for Equity in Health*, *19*(1). https://doi.org/10.1186/s12939-020-01241-0

Kusters, R., Misevic, D., Berry, H., Cully, A., Le Cunff, Y., Dandoy, L., Díaz-Rodríguez, N., Ficher, M., Grizou, J., Othmani, A., Palpanas, T., Komorowski, M., Loiseau, P., Moulin Frier, C., Nanini, S., Quercia, D., Sebag, M., Soulié Fogelman, F., Taleb, S., … Wehbi, F. (2020). Interdisciplinary research in artificial intelligence: Challenges and opportunities. *Frontiers in Big Data*, *3*. https://doi.org/10.3389/fdata.2020.577974

Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., Malhotra, N., Lui, V., & Gibson, J. (2021). Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Medical Ethics*, *22*(1). https://doi.org/10.1186/s12910-021-00577-8

Nguyen, V. T., Engleton, M., Davison, M., Ravaud, P., Porcher, R., & Boutron, I. (2021). Risk of bias in observational studies using routinely collected data of comparative effectiveness research: a meta-research study. *BMC Medicine*, *19*(1). https://doi.org/10.1186/s12916-021-02151-w

Ruprecht, M. M., Wang, X., Johnson, A. K., Xu, J., Felt, D., Ihenacho, S., Stonehouse, P., Curry, C. W., DeBroux, C., Costa, D., & Phillips, G., II. (2021). Evidence of social and structural COVID-19 disparities by sexual orientation, gender identity, and race/ethnicity in an urban environment. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, *98*(1), 27–40.

Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01488-9

STANDING together - 1 - documentation recommendations. (n.d.). Retrieved October 27, 2023, from https://www.datadiversity.org/draft-standards/1-documentation-recommendations

Stone, K., Zwiggelaar, R., Jones, P., & Mac Parthaláin, N. (2022). A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health*, *1*(4), e0000017.

Veenstra, G. (2011). Race, gender, class, and sexual orientation: intersecting axes of inequality and self-rated health in Canada. *International Journal for Equity in Health*, *10*(1), 3.

Wilkinson, M. D., Dumontier, M., Jan Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2019). Addendum: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *6*(1). https://doi.org/10.1038/s41597-019-0009-6

Xu, Z., Zhao, C., Scales, C. D., Jr, Henao, R., & Goldstein, B. A. (2022). Predicting in-hospital length of stay: a two-stage modeling approach to account for highly skewed data. *BMC Medical Informatics and Decision Making*, *22*(1). https://doi.org/10.1186/s12911-022-01855-0

Zimlichman Eyal, Nicklin Wendy, Aggarwal Rajesh, Bates David W. (2021). Health Care 2030: The Coming Transformation. *Catalyst Nej*. https://doi.org/10.1056/CAT.20.0569

Table 1. Description of the population contained in the dataset

		Overall	**Missing**
n		43154	
sex (%)	Female	21379 (49.5)	0

	Male	21775 (50.5)	
Age mean (SD)		58.4 (23.8)	0
Hospitalization **, n (%)**	Intermediate Care care	547 (1.3)	
	Adult Hospitalization	34815 (80.7)	
	Adult ICU	1571 (3.6)	
	Emergency	5571 (12.9)	
	Pediatric Emergency	620 (1.4)	
Type of Hospitalization **, n (%)**	Adult Intensive Care	1573 (3.7)	209
	Intermediate Care Adults	548 (1.3)	
	General Wards Adults	32882 (76.6)	
	General Paediatric Wards	1928 (4.5)	
	No information	6 (0.0)	
	24<hour emergencies	425 (1.0)	
	Emergencies >= 24 hours	5583 (13.0)	
Life cycle, **n (%)**	Adolescence	869 (2.0)	0
	Adults	11555 (26.8)	

	Infancy	657 (1.5)
	Youth	3257 (7.5)
	Early childhood	1430 (3.3)
	Geriatric	25386 (58.8)

Speciality, **n (%)**	Allergology	1 (0.0)	0

Anesthesiology	1 (0.0)
Cardiology	2 (0.0)
Cardiovascular Surgery	463 (1.1)
General Surgery	5780 (13.4)
Hepatobiliopancreatic surgery	755 (1.7)
Maxillofacial Surgery	595 (1.4)
Oncological Surgery	107 (0.2)
Plastic- Maxillofacial and Hand Surgery	1378 (3.2)
Thoracic Surgery	844 (2.0)
Vascular Surgery	1776 (4.1)
Head and Neck Surgery	575 (1.3)
Transplant Surgery	15 (0.0)
Epileptology	3 (0.0)
Gynaecological Oncology	123 (0.3)
Gynaecology	220 (0.5)
Haematology	960 (2.2)
Hematology Transplant	13 (0.0)
Hepatology	582 (1.3)
Hepatology Transplantation	21 (0.0)
Infectious Diseases	1011 (2.3)
Intensivist	498 (1.2)
General Medicine	891 (2.1)
Internal Medicine	8975 (20.8)
Nephrology	951 (2.2)
Pediatric Nephrology	2 (0.0)
Nephrology Transplantation	139 (0.3)
Neurosurgery	2135 (4.9)
Neurology	2338 (5.4)
Paediatric Neurology	4 (0.0)
Ophthalmology	23 (0.1)

	Oncology	635 (1.5)	
	Orthopaedics	4519 (10.5)	
	Orthopedic Oncology	4 (0.0)	
	Otorhinolaryngology	393 (0.9)	
	Paediatrics	2114 (4.9)	
	Psychiatry	548 (1.3)	
	Toxicology	380 (0.9)	
	Urgencyology	265 (0.6)	
	Urology	2991 (6.9)	
	Vascular **Medicine**	120 (0.3)	
Discharge Hours, **mean (SD)**		49.7 (78.5)	0
Days of Clinical Stay, **mean (SD)**		8.8 (10.9)	0
Modality of admission **n (%)**	Outpatient or Scheduled Consultation	3289 (7.6)	0
	Born in the Institution	2 (0.0)	
	Referred	14017 (32.5)	
	No information	11 (0.0)	
	Emergency	25835 (59.9)	
Type of Discharge **, n (%)**	Medical Discharge	37774 (87.5)	0
	Voluntary Discharge	355 (0.8)	
	Mortality	3367 (7.8)	
	Escape	35 (0.1)	
	Referral	1617 (3.7)	

	No information	6 (0.0)