

Towards Patient-Driven Phenotyping and Similarity for Precision Medicine

Tiffany J. Callahan, MPH¹, Olivier Bodenreider, MD, PhD², Michael G. Kahn, MD, PhD³

¹Computational Bioscience Program, University of Colorado Denver Anschutz Medical Campus, Aurora, CO; ²Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD; ³Department of Pediatrics, University of Colorado Denver Anschutz Medical Campus, Aurora CO

Abstract

Clinical phenotyping provides important insight into the manifestation and outcome of rare and complex diseases. Traditional phenotyping techniques often require multiple iterations of refinement with a domain expert, lack interoperability, and have limited reproducibility. In comparison, patient similarity-based techniques derive personalized patient risk models that are highly accurate, even when applied to sparse data or poorly characterized diseases/outcomes. We present preliminary results from a novel, unsupervised data-driven method for applying patient similarity to pediatric phenotyping.

Introduction

Clinical phenotyping, or the classification of patients with or without an outcome or disease, is a technique designed to provide clinicians with important insight into the development, progression, and outcome of complex diseases for a population of patients. While there is a large body of research supporting the successful implementation of existing phenotyping methods, these techniques often require multiple iterations of refinement, lack generalizability and interoperability, and have limited reproducibility.¹⁻³ Compared to traditional phenotyping approaches, patient similarity-based techniques aim to derive personalized risk models for each patient. When compared to other methods, patient similarity-based approaches have shown to be more accurate,⁴ even when applied to sparse data or poorly characterized diseases/phenotypes.⁵ An example of clinician driven supervised patient-similarity-based methods is Longhurst and Shah's "Green Button".⁶ While this approach, and others like it, are scalable and accurate, they still suffer from poor handling of missing data, lack robust internal and external validation, and maintain reliance on domain expertise.⁷ The current project aims to address some of the limitations of traditional phenotyping and existing expert-driven patient similarity-based methods by developing a novel, unsupervised data-driven algorithm for patient-level phenotyping and similarity.

Methods

A composite patient similarity algorithm was designed specifically for use with the Observational Medical Outcomes Partnership (OMOP) common data model (CDM).⁸ By developing our method specifically for this CDM, we can take advantage of pre-normalized data standardized to a specific set of clinical terminologies and can provide a tool that can be readily adopted by members of the worldwide OMOP community. Our composite patient similarity algorithm leverages existing pairwise⁹ and groupwise¹⁰ semantic similarity measures. Pairwise similarity scores were calculated for demographic attributes, accounting for binary (e.g., gender), categorical (e.g., race), and continuous (e.g., age) variables. Pairwise similarity scores were calculated for clinical attributes by incorporating hierarchical relations from standard clinical terminologies (e.g., LOINC, RxNorm, and SNOMED CT). Furthermore, we used groupwise similarity measures to compare sets of codes among patients. The final composite similarity score, where scores range from 0.0 (completely dissimilar) to 1.0 (perfect similarity), between two patients is calculated as a weighted average of the individual demographic and clinical groupwise similarities. While individual attribute weights can be learned, or user generated, no differential weighting was applied in this experiment.

A proof-of-concept demonstration of the composite patient similarity algorithm was performed using de-identified Children's Hospital of Colorado (CHCO). CHCO data conforms to the structure defined by the PEDSnet, which is an adaptation of the OMOP CDM version 5.0.^{8,11} From the condition occurrence, drug exposure, measurement, observation, and procedures tables, we retrieved demographic and clinical data and constructed two distinct groups of 10 patients having the highest counts of cystic fibrosis (CF; SNOMED CT 190905008) and Huntington's Chorea (HC; SNOMED CT 58756001) encounter-diagnoses. To ensure an unbiased assessment of the method, all SNOMED CT codes for CF and HC used to define the two groups were excluded. Agglomerative hierarchical clustering with complete linkage and Euclidean distance were used to generate clusters of similar patients in the expectation that the

two groups of patients would separate into distinct clusters. Results were described and interpreted using dendrograms and heat maps. This project was approved by the Colorado Multiple Institutional Review Board (15-0445).

Results

Patients were predominately white (90%) and female (60%) with a median age of 19. Hierarchical clustering resulted in four groups of semantically similar patients with scores ranging from 0.36 to 1.0 (Figure 1, <https://tinyurl.com/y6uzlx6u>). The red (n=3) and blue (n=2) clusters only contained HC patients. On average, HC red cluster patients were younger (17 vs. 26 years) than blue cluster HC patients. Red cluster patients were distinguished by more frequent Parkinson's disease (16.39%), dystonia (12.61%), and failure to thrive (7.56%) encounter-diagnoses. There were no occurrences of these diagnosis codes among any of the blue cluster HC patients. Further, medical nutrition therapy, which occurred in only two encounters in blue cluster HC patients, was the only frequently co-occurring red and blue cluster HC patient encounter-procedure. The white cluster (n=9) only contained CF patients. Headache (7.50%), anxiety disorder (6.80%), and asthma (5.81%) were the most frequent encounter-diagnoses. Pressurized or nonpressurized inhalation treatment for acute airway obstruction (18.05%), manipulation of chest wall to facilitate lung function (9.99%), and demonstration/evaluation of patient utilization of an aerosol generator (7.68%) were the most frequent encounter-procedures. The final magenta cluster (n=6) contained 5 HC patients and 1 CF patient. These patients were most frequently diagnosed with post inflammatory pulmonary fibrosis (6.51%), hypoxemia (5.41%), and congenital iodine deficiency (4.31%). Their most frequent encounter-procedures were noninvasive ear/pulse oximetry for oxygen saturation (8.68%), pressurized or nonpressurized inhalation treatment for acute airway obstruction (7.47%), and collection of venous blood by venipuncture (5.45%). A detailed description and comparison of the patient's clinical attributes, by cluster, is provided in Figures 1-2 (<https://tinyurl.com/y6uzlx6u>).

Discussion

We are currently developing a novel unsupervised data-driven method to measure patient similarity and provided an initial proof-of-concept using a sample of pediatric patients. Preliminary results highlight the ability of our approach to successfully identify clinically distinguishable groups and sub-groups of similar patients, in the absence of the patient's primary diagnoses. Future work is underway to address current limitations including: conducting a more comprehensive evaluation, accounting for changes in clinical variables over time, and learning of variable weights.

References

1. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc JAMIA*. 2014 Apr;21(2):221–30.
2. Richesson RL, Smerek MM, Blake Cameron C. A Framework to Support the Sharing and Reuse of Computable Phenotype Definitions Across Health Care Delivery and Clinical Research Applications. *eGEMs [Internet]*. 2016 Jul 5;4(3). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975566/>
3. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med*. 2016 Jul;71:57–61.
4. Ng K, Sun J, Hu J, Wang F. Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity. *AMIA Summits Transl Sci Proc*. 2015 Mar 25;2015:132–6.
5. Beaulieu-Jones BK, Greene CS, Pooled Resource Open-Access ALS Clinical Trials Consortium. Semi-supervised learning of the electronic health record for phenotype stratification. *J Biomed Inform*. 2016 Dec;64:168–78.
6. Longhurst CA, Harrington RA, Shah NH. A “Green Button” For Using Aggregate Patient Data At The Point Of Care. *Health Aff (Millwood)*. 2014 Jul 1;33(7):1229–35.
7. Sharafoddini A, Dubin JA, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. *JMIR Med Inform*. 2017;5(1):e7.
8. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc JAMIA*. 2012;19(1):54–60.
9. Batet M, Sánchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. *J Biomed Inform*. 2011 Feb;44(1):118–25.
10. Azuaje F, Wang H, Bodenreider O. Ontology-driven similarity approaches to supporting gene functional assessment. In: *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies*. 2005. p. 9–10.
11. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc JAMIA*. 2014 Jul;21(4):602–6.

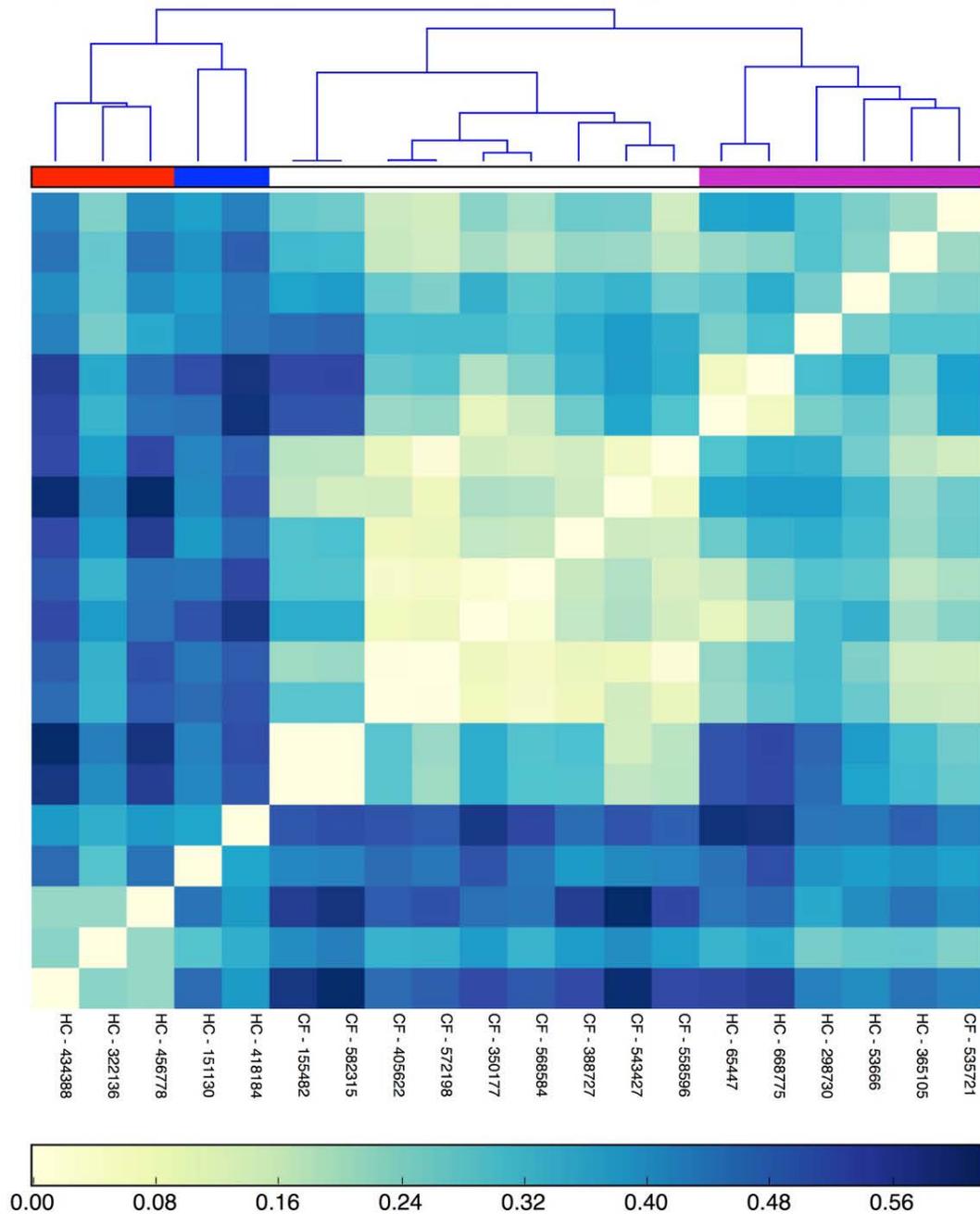


Figure 1. Patient similarity scores using age and encounter-diagnoses, measurements, encounter-medications, observations, and encounter-procedures ranged from 0.36-1.0. An intra-cluster description for each of the four identified clusters is provided below. Note that the SNOMED CT codes used to identify CF and HC patients were excluded when determining the most frequently occurring encounter-diagnoses. Additionally, “office or other basic outpatient visit” level of service procedure codes were also excluded when calculating the most frequently occurring encounter-procedures. Due to space limitations, we limit our interpretation to diagnoses, medications, and procedures.

Red Cluster Patients. The red cluster (n=3) contained only HC patients. The most frequent encounter-diagnoses included: Parkinson’s (16.39%), dystonia (12.61%), and failure to thrive (7.56%). The most frequent encounter-medications included: carbidopa-levodopa 25-100 MG (24.24%), carbidopa 25 MG (18.18%), and acetaminophen 80 MG (9.10%). The most frequent encounter-procedures included: medical nutrition therapy (13.04%), motion fluoroscopic evaluation (6.52%), and evaluation of oral and pharyngeal swallowing function (6.52%).

Blue Cluster Patients. The blue cluster (n=2) contained only HC patients. The most frequent encounter-diagnoses included: mixed receptive-expressive language disorder (19.45%), incoordination (13.65%), and child attention deficit disorder (12.97%). The most frequent encounter-medications included: choral hydrate 500 MG (16.67%), celexa 10 MG (16.67%), and lupron depot-ped 11.25 MG (12.50%). The most frequent encounter-procedures included: neuromuscular reeducation of movements (21.01%), dynamic activities to improve functional performance (19.70%), and therapeutic exercises to develop strength and endurance (11.59%).

White Cluster Patients. The white cluster (n=9) contained only CF patients. The most frequent encounter-diagnoses included: headache (7.50%), anxiety disorder (6.80%), and asthma (5.81%). The most frequent encounter-medications included: dornase alfa 1 MG (5.55%), lansoprazole 30 MG (3.00%), and fluticasone propionate 50 MCG (2.64%). The most frequent encounter-procedures included: pressurized or nonpressurized inhalation treatment for acute airway obstruction or for sputum induction for diagnostic purposes (18.05%), manipulation chest wall, such as cupping, percussing, and vibration to facilitate lung function; subsequent (9.99%), and demonstration and/or evaluation of patient utilization of an aerosol generator, nebulizer, metered dose inhaler or IPPB device (7.68%).

Magenta Cluster Patients. The magenta cluster (n=6) contained 5 HC patients and 1 CF patient. The most frequent encounter-diagnoses included: post inflammatory pulmonary fibrosis (6.51%), hypoxemia (5.41%), and congenital iodine deficiency (4.31%). The most frequent encounter-medications included: albuterol sulfate 2.5 MG (9.30%), acetaminophen 80 MG (6.03%), and cefuroxime 30 MG (4.08%). The most frequent encounter-procedures included: noninvasive ear or pulse oximetry for oxygen saturation; by continuous overnight monitoring (8.68%), pressurized or nonpressurized inhalation treatment for acute airway obstruction or for sputum induction for diagnostic purposes (7.47%), and collection of venous blood by venipuncture (5.45%).

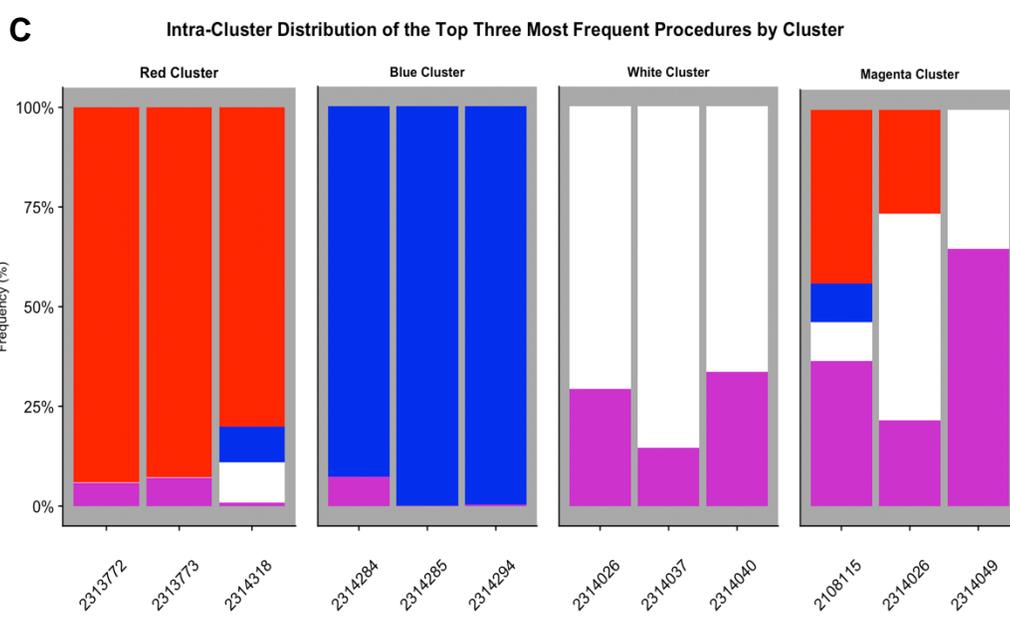
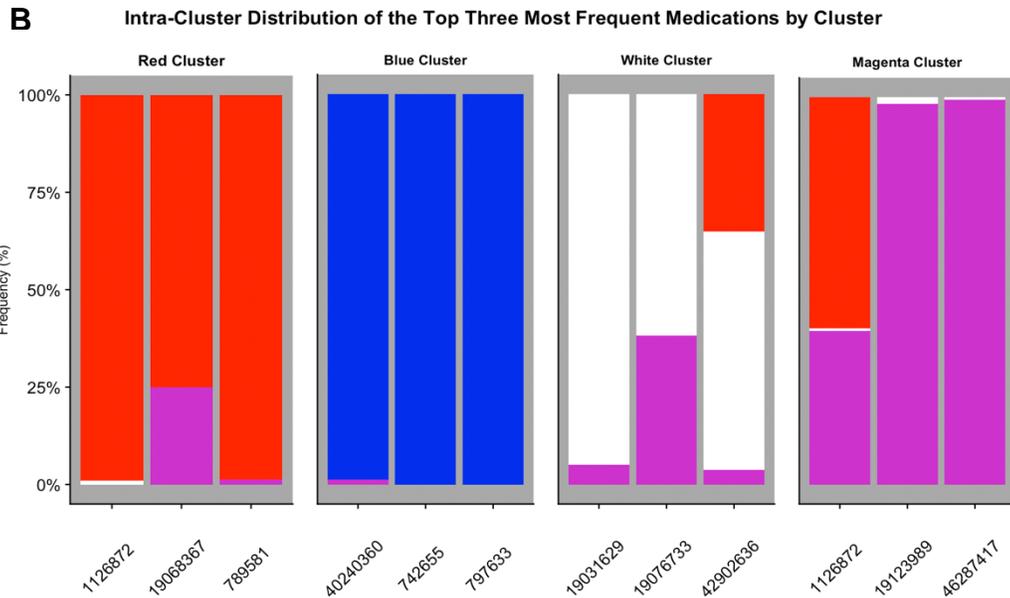
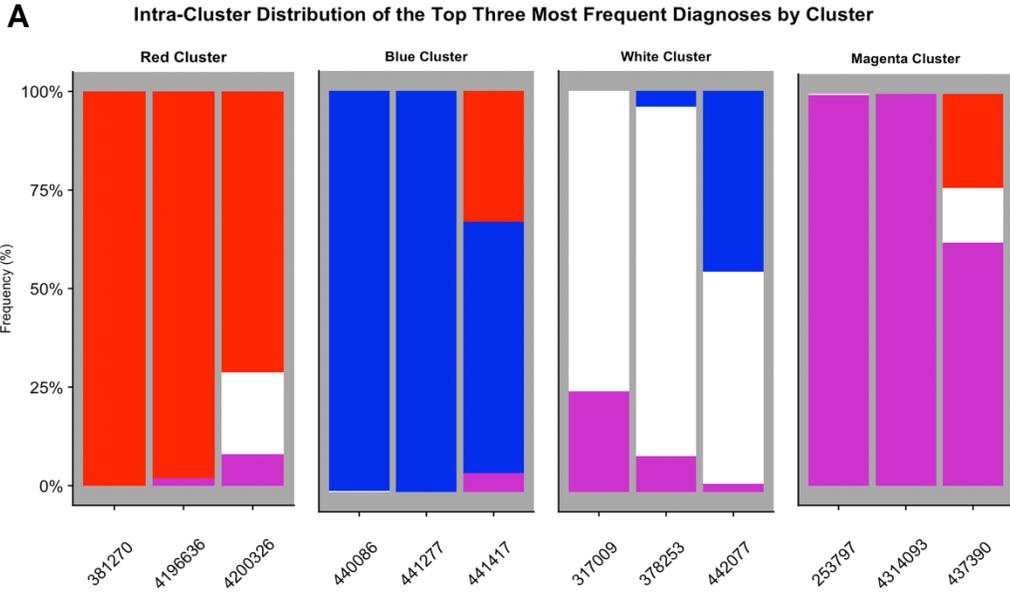


Figure 2. These figures provide a between-cluster comparison of the three most frequently occurring encounter-diagnoses (A), encounter-medications (B), and encounter-procedures (C). Each cluster is colored to be consistent with Figure 1. Within each of the figures, there is a stacked-bar chart for each of the four “primary” clusters from Figure 1 (i.e., red, blue, white, and magenta). These stacked-bar charts represent the distribution of a cluster’s three most frequently occurring encounter-diagnoses, medication or encounter-procedures (shown as OMOP concept identifiers on the x-axis) codes in comparison to the other clusters. For example, the first stacked-bar plot in Figure 2A represents the frequency of the three most frequently red cluster patient encounter-diagnoses. The absence of any blue, white, or magenta on the first bar in this chart means that none of the patients in those clusters had any occurrence of that specific diagnosis code. In comparison, the third bar in that chart, shown as primarily red with some white and magenta color, suggests that while no patients from the blue cluster had any occurrence of that specific diagnosis code, a few occurrences of the diagnosis code occurred for patients in the white and magenta clusters. In general, these figures provide evidence suggesting that there is little overlap between the most frequently occurring encounter-diagnoses, encounter-medications, and encounter-procedures among the patients in the red, blue, and white clusters. A description of each OMOP code is provided in the table on page 3.

CODE	CODE DESCRIPTION
381270	Parkinson's disease
4196636	Dystonia
4200326	Failure to Thrive in infant
441277	Mixed receptive-expressive language disorder
441417	Incoordination
440086	Child attention deficit disorder
378253	headache
442077	Anxiety disorder
317009	Asthma
253797	Post-inflammatory pulmonary fibrosis
437390	Hypoxemia
4314093	Congenital iodine deficiency syndrome
789581	CARBIDOPA-LEVODOPA 25-100 MG PO TABS
19068367	CARBIDOPA 25 MG PO TABS
1126872	ACETAMINOPHEN 80 MG/0.8ML PO SUSPENSION
742655	CHLORAL HYDRATE 500 MG/5ML PO SYRUP
797633	CELEXA 10 MG PO TABS
40240360	LUPRON DEPOT-PED 11.25 MG IM KIT
19076733	DORNASE ALFA 1 MG/ML INH SOLUTION
19031629	LANSOPRAZOLE 30 MG PO DR CAPSULE
42902636	FLUTICASONE PROPIONATE 50 MCG/ACT NASAL SUSPENSION
19123989	ALBUTEROL SULFATE (2.5 MG/3ML) 0.083% INH NEB SOLN
46287417	CEFUROXIME (30 MG/ML) IV
2314318	Medical nutrition therapy; initial assessment and intervention, individual, face-to-face with the patient, each 15 minutes
2313773	Motion fluoroscopic evaluation of swallowing function by cine or video recording
2313772	Evaluation of oral and pharyngeal swallowing function
2314285	Therapeutic procedure, 1 or more areas, each 15 minutes; neuromuscular reeducation of movement, balance, coordination, kinesthetic sense, posture, and/or proprioception for sitting and/or standing activities
2314294	Therapeutic activities, direct (one-on-one) patient contact (use of dynamic activities to improve functional performance), each 15 minutes
2314284	Therapeutic procedure, 1 or more areas, each 15 minutes; therapeutic exercises to develop strength and endurance, range of motion and flexibility
2314026	Pressurized or nonpressurized inhalation treatment for acute airway obstruction or for sputum induction for diagnostic purposes (eg, with an aerosol generator, nebulizer, metered dose inhaler or intermittent positive pressure breathing [IPPB] device)
2314040	Manipulation chest wall, such as cupping, percussing, and vibration to facilitate lung function; subsequent
2314037	Demonstration and/or evaluation of patient utilization of an aerosol generator, nebulizer, metered dose inhaler or IPPB device
2314049	Noninvasive ear or pulse oximetry for oxygen saturation; by continuous overnight monitoring (separate procedure)
2108115	Collection of venous blood by venipuncture