大規模言語モデルを用いた看護 OSCE の自動採点の検証

Validation of Automatic Scoring of Nursing OSCE using A Large Language Models

佐藤 玲央[*1], 上野 春毅[*2], 塚本 容子[*3], 小松川 浩[*1]
Reo SATOU[*1], Haruki UENO[*2], Yoko TSUKAMOTO[*3], Hiroshi KOMATSUGAWA[*1]
[*1] 公立千歳科学技術大学大学院 理工学研究科
[*1]Graduate School of Science and Technology, Chitose Institute of Science and Technology
[*2] 公立千歳科学技術大学 理工学部
[*2]Faculty of Science and Technology, Chitose Institute of Science and Technology
[*3] 北海道医療大学大学院 看護福祉学研究科
[*3]School of Nursing & Social Services Health Sciences University of Hokkaido
Email: m2230170@photon.chitose.ac.jp

あらまし：本研究は，ChatGPT が指導者の評価を代替する可能性を検証することを目的に OSCE（客観的臨床能力試験）を自動採点した結果と指導者がつけた評価との一致率と採点結果の分布の偏りを探索的に分析した．結果，自動採点は中間的評価では一致率が高い一方，端に位置する評価の一致率は低く，指導者の評価全てを代替することは難しいが，プロンプトの改良で精度向上の可能性が示された．
キーワード：看護 OSCE，看護学教育，LLM，ChatGPT，GPT-4o

1. はじめに

看護師養成機関では，実際の現場が求める問題解決力を基礎とした看護的思考能力を備える人材の育成を，看護基礎教育段階において試みている．看護的思考能力を座学での事例展開と臨地実習を繰り返すことを通じて身につけていく．この過程において，主体的・能動的な問題解決に重要な臨床場面をいかに現実に近づけ具体化し，看護的思考能力を養成できるかが大きな課題となってきた．

こうした課題に対し，本研究チームでは先行研究において看護的思考能力の向上を狙い，事例展開学習における看護実習のシナリオに則り，看護的思考過程を自己トレーニングできる臨床判断支援システムを提案し検証してきた(1)．しかし，臨地実習を視野に入れると，実際の現場での行動，発言，記録などの様々な情報を統合した評価とフィードバックが求められる．

一方，大規模言語モデル（Large Language Models; LLM）の性能が急速に進展しており，様々な産業分野での活用が期待されている．寺岡の研究では看護学教育分野において LLM の一種である ChatGPT が教育支援ツールになり得ることが示唆されている(2)．

本研究では，ペーパーテスト等で測定が困難な技能等の精神運動領域や態度・習慣等の情意領域を含めた臨地実習を通して習得された看護実践能力を評価する客観的臨床能力試験（Objective Structured Clinical Examination：OSCE）を自動的に採点して評価結果を学生にフィードバックする手法の確立を目指し，指導者の評価の代替可能性を検証する．

2. 研究の目的

OSCE を ChatGPT で自動採点した結果と指導者がつけた評価との一致率と採点結果の分布の偏りを分析し，指導者の評価を代替する可能性を検証することを目的とする．

3. OSCE の自動採点方法

OSCE では実習生が患者にヒアリングした情報や活動時に知り得た情報を適宜記録し，記録を参照しながら発言や判断を行うことを想定する．本研究では，指導者の採点が実習生の記録と関係があると仮定し，OSCE を受けた実習生の記録と OSCE 評価表を基に ChatGPT で自動採点する方法をとる．この際に，自動採点した結果と指導者がつけた評価との一致率をみる．

3.1 OSCE 評価表の評価項目

今回対象とする評価項目は 32 項目である．大枠では 5 種類に分類され，全体を通しての評価，医療面接の評価，フィジカルイグザミネーション，患者への説明，診断である．全体を通しての評価は 10 項目ある．例として，患者の本人確認ができることを評価項目とし，姓名・生年月日・住所を患者自身から発言を引き出しているかを評価のポイントとする．医療面接の評価は 10 項目ある．例として，排尿障害の経時的変化の確認を評価項目とし，3 日前から始まった排尿時痛，頻回の尿意を確認できているかを評価のポイントとする．他項目にも評価項目と評価のポイントが設定されている．これら 32 項目それぞれについて A:できた，B:概ねできた，C:あまりできなかった，D:できなかった，以上 4 段階の評価基準で採点する．

3.2 自動採点の方法

OpenAI 社が提供する ChatGPT の GPT-4o のモデルを活用する．実習生が記録したデータを質問メッセージと見立て，採点結果を出力するように指示した．

この際のプロンプトの構成は，GPT 自身の役割，OSCE のシナリオ患者の情報，OSCE 評価表に記載されている評価基準，出力フォーマット，実習生の記録とした．このとき OSCE 評価表の 32 項目ある評価項目を一括で入力する場合（手法 1）と 1 項目ずつ入力する場合（手法 2）を実施し，プロンプトの構成による一致率の変化を比較した．表 1 と表 2 に簡略化したプロンプトを示す．

表 1 設計したプロンプトの例（手法 1）

プロンプト
役割 ### あなた(GPT)の役割は「客観的臨床能力試験の試験監督」です。 ### シナリオ患者 ### 52歳、女性、受診理由:(イライラした様子で入室してくる)… ### 評価基準 ### [番号・全般を通しての評価・評価のポイント] [1・患者の本人確認ができる・姓名、生年月日、住所を患者自身から言ってもらう] [2・挨拶、自己紹介が出来る・受験者の名前、所属を述べる] ~~~~~手法1の場合、32項目を一括入力~~~~~ ### フォーマット ### 【評価】:~~ここにA~Dの評価を記載~~ 【コメント】:~~ここに評価の説明を記載~~ ### 学生の記録 ### <CC>一日に何回もトイレに行って、イライラする。

表 2 設計したプロンプトの例（手法 2）

プロンプト
~~以上、手法1と共通~~~~~ ### 評価基準 ### [番号・全般を通しての評価・評価のポイント] [1・患者の本人確認ができる・姓名、生年月日、住所を患者自身から言ってもらう] ~~~~~手法2の場合、項目毎にプロンプトを入れえる~~~~ ~~~~~以下、手法1と共通~~~~~

4. 検証・結果

北海道医療大学から関係者に同意のもと OSCE に基づいて採点された実習生の記録と指導者の採点結果のデータを個人の特定ができない形式に変換した上で 6 名分の提供を受けた．

手法 1 と手法 2 における指導者評価と自動採点評価の一致率と 4 段階の各評価についての F 値を表 3 に示す．手法 2 は手法 1 と比べて全ての数値が改善しており，1 項目毎に評価を出力する方が精度が高いことが確認された．この結果は手法 2 の方がプロンプトの長さが短いことに起因すると考えられる．

表 4 に手法 2 の場合における 6 名分全 32 項目の採点結果の分布を示す．指導者・自動採点評価共に B 又は C の中間的な評価に分布している傾向がやや強くなっている．B 及び C 評価の F 値, 41.43%と 41.32%に対し A 及び D 評価の F 値は 2.94%と 14.55%となっていることから中間的な評価は一定程度一致するが，端に位置する評価は一致しないと言える．

表 5, 表 6, 表 7 に手法 2 における項目の大枠での分類毎の採点結果の分類を示す．表 5 の全体を通しての評価の項目は指導者が A 又は C 評価に対し，自動採点が C 及び D 評価に多く分布しており自動採点が厳しい評価をする傾向にあると言える．表 6 の医療面接の評価の項目は指導者が B から D 評価に多く分布し，自動採点が A から C 評価に多く分布することから，やや自動採点が易しい評価をする傾向にあると言える．表 7 の医療面接の評価の項目は指導者・自動採点ともに A から D 評価に広く分散して分布しており評価の厳しさに偏りは見られない. なお，患者への説明（項目 29-30），診断（項目 31-32）はデータ数が少ないため割愛した．

表 3 評価の一致率と F 値

	一致率	F値-A評価	F値-B評価	F値-C評価	F値-D評価
手法1	23.44%	0.00%	35.62%	31.58%	4.17%
手法2	30.73%	2.94%	41.43%	41.32%	14.55%

表 4 全 32 項目の採点結果

全32項目	指導者評価				
自動採点評価	A	B	C	D	小計
A	1	12	7	6	26
B	8	29	26	5	68
C	12	21	25	3	61
D	21	10	2	4	37
小計	42	72	60	18	192

表 5 全体を通しての評価の採点結果

項目1-10	指導者評価				
自動採点評価	A	B	C	D	小計
A	1	0	0	0	1
B	8	0	10	0	18
C	8	0	14	0	22
D	19	0	0	0	19
小計	36	0	24	0	60

表 6 医療面接の評価の採点結果

項目11-20	指導者評価				
自動採点評価	A	B	C	D	小計
A	0	8	3	3	14
B	0	20	4	2	26
C	0	13	5	1	19
D	0	1	0	0	1
小計	0	42	12	6	60

表 7 フィジカルイグザミネーションの採点結果

項目21-28	指導者評価				
自動採点評価	A	B	C	D	小計
A	0	4	4	3	11
B	0	5	4	3	12
C	4	5	2	2	13
D	2	4	2	4	12
小計	6	18	12	12	48

5. おわりに

OSCE を ChatGPT で自動採点した結果と指導者がつけた評価との一致率と採点結果の分布の偏りを探索的に分析した．その結果，プロンプトの改善による自動採点の精度向上が見込めると共に，端に位置する評価を正確に採点することは中間的評価に比べて難しく，評価項目により採点の厳しさに偏りが生じることが確認された．今回確認された傾向を踏まえると指導者の評価の全てを代替する可能性は低いと考えられるが，現在 OSCE 評価表の項目をより細分化し，項目単位で採点の代替可能性を分析している．

参考文献

(1) 佐藤玲央: “看護過程での臨床判断支援システムにおける大規模言語モデルを活用したフィードバックの自動生成”, JSISE 研究会研究報告, Vol.38, No.2, pp.14-18（2023）
(2) 寺岡三左子: “ChatGPT®は有効な教育支援ツールになり得るか？-看護学における対話型 AI をめぐる議論の動向-”, 医療看護研究, Vol.20, No.1, pp.64-70 （2023）