

医療従事者向け対話訓練システムのための 各種精神疾患特性を持つ Chatbot の構築

水田 龍汰[†] 小谷 涼太[†] Siriaraya Panote^{††} She Wan-Jou^{††} Cheng Asaka Lan[†]
中島 伸介[†]

[†] 京都産業大学情報理工学部情報理工学科 〒603-8555 京都府京都市北区上賀茂本山

^{††} 京都工芸繊維大学 情報工学・人間科学系 〒606-8585 京都府京都市左京区松ヶ崎橋上町

E-mail: [†]{g2354521,g2253523,a007532,nakajima}@cc.kyoto-su.ac.jp, ^{††}{spanote,wjs2004}@kit.ac.jp

あらまし 近年、精神疾患の患者数は増加傾向にあり、医学生や医療従事者が精神疾患の特徴に対する理解や適切なコミュニケーションスキルの習得がますます重要となっている。しかし、精神疾患医療で活用されてきた模擬患者による演習は高い人的コストや準備負担が生じる。そこで本研究では、大規模言語モデルを用いた医療従事者向け対話トレーニングシステムを提案する。本システムでは、大規模言語モデルを用いて、双極性障害（躁鬱）、PTSD、アルコール依存症などの特定の精神疾患の特徴や症状、会話パターンを ChatBot に付与する。これにより、ユーザ（医療従事者や学生）は、模擬患者としての ChatBot を相手に、診断に必要な情報収集や信頼関係形成のための対話を模擬的に練習することが可能となる。

キーワード ヘルスケア, LLM, 対話システム, 模擬患者

1 はじめに

近年、新型コロナウイルス感染症（COVID-19）の世界的な拡大や SNS の普及により、うつ病等の精神疾患を有する患者数が増加傾向にある [1]。これにより、医療従事者が精神疾患の特徴に対する理解が重要になってきている。そして、医療従事者と患者との間のコミュニケーションは、適切な診療を行う上で不可欠である。特に精神疾患医療では、患者との対話を通じて信頼関係の構築と情報収集が重要となる。そのため、医療従事者は精神疾患における模擬患者との対話訓練を行うことで、患者とのコミュニケーションスキルを学ぶ。

模擬患者とは 1964 年に米国の Barrows がその利用を報告し、日本へは 1975 年に同氏によって紹介された教育手法であり、医学・薬学教育において患者役を演じるために訓練された人々のことをいう [2]。模擬患者を用いた実習を受けることで、学生同士のロールプレイでは得られない「臨場感」や「緊張感」を持って医療面接や服薬指導を体験できることや、模擬患者からの市民感覚に基づいたフィードバックを通じて、患者の気持ちに配慮する態度やコミュニケーション能力の向上が期待できる。

しかし、模擬患者の導入や運営には課題が存在する。藤崎ら [3] は、模擬患者活動の拡大に伴い、模擬患者の演技の質や教育効果を維持することが課題となっていると指摘している。特に、模擬患者を適切に指導・運営するための指導者養成が十分に確立されていない点が問題とされている。また、模擬患者の多くが市民ボランティアであることから、謝金の支払いが大学等の制度上想定されておらず、その結果として資金面・運営面での課題が生じていることが示されている。このような背景

から、従来の模擬患者演習に代わる、より実施しやすい対話訓練手法の構築が求められている。

そこで本研究の最終目標は、医療従事者が場所や時間を問わず、精神疾患患者との対話を想定した訓練を行える対話トレーニングシステムを開発することである。本システムでは、実際の診療場面を想定した対話練習環境を提供するため、Chatbot を模擬患者として用いる。これにより、ユーザは診断に必要な情報収集や患者との信頼関係構築を意識した対話を、実際の患者対応を想定して繰り返し練習することが可能となる。また、本システムは場所や時間にとらわれず対話訓練を提供できるため、従来の対面型訓練における人的コストや時間的制約といった課題を軽減し、医療従事者が柔軟かつ継続的に対話能力を向上させることが期待される。本稿では、この最終目標に到達するための第一段階として、各種精神疾患特性を有する Chatbot の構築を行う。

2 関連研究

本田らは看護学生を対象とした模擬患者を用いた演習の効果を検証し、模擬患者との対話を通じて、学生が「安心感を与える環境づくり」や「共感・傾聴的態度」の重要性を認識し、対象者のニーズを明確にするための基本的対話技術を学習できたと示されている。また、模擬患者からのフィードバックは、学生にとって「本物の患者以上のリアリティ」を感じさせ、自身の態度や言葉遣いの不足に気づく貴重な機会となることが示されている [4]。

Yamamoto らは本の医学部 4 年生を対象に、大規模言語モデルを用いた AI 模擬患者システムによる介入研究を行っている。

同研究では、AI 模擬患者を用いて自習を行った学生群は、従来の教育のみを受けた対照群と比較して、全国共用試験における医療面接のスコアが有意に高かったことが示された。この研究により、AI を用いたトレーニングシステムは安全であり、学生が場所や時間を選ばずに繰り返し練習することで自己効力感を高め、一定の教育効果をもたらすことが実証されている [5]。

遠山らは地方における模擬患者確保の困難さという医療面接教育上の課題に着目し、大規模言語モデル（GPT-4o-mini）を用いた AI 模擬患者モデルの構築手法を提案している。疾患シナリオから医療面接における質問と回答の対話データを自動生成する学習データ作成 AI と、そのデータを用いてファインチューニングされた模擬患者 AI から成る 2 段階のモデル構成を採用している。評価実験の結果、提案手法は従来の LLM に見られたハルシネーションや過剰回答を抑制し、適切な情報量で応答可能であることが示された。また、医師による評価においても、模擬患者としての会話能力および実用性に関して高い評価が得られており、AI を用いた模擬患者システムが医療面接教育において有効な訓練手法となり得ることが示唆されている [6]。

Lee らは大規模言語モデルを用いて複数の精神疾患を体系的にシミュレートするフレームワーク「PSYPHE」を提案している。PSYPHE は、双極性障害、PTSD、パニック障害を含む計 7 種類の精神疾患を対象とし、各疾患に特有の症状表現や対話傾向を LLM に付与することで、精神疾患患者としての応答を生成する枠組みである。また、精神科医による専門的評価として「Conformity Score（適合性スコア）」を導入し、生成された応答が各疾患の臨床的特徴とどの程度整合しているかを定量的に評価している。その結果、PSYPHE によって生成された対話は、各精神疾患の特徴を高い精度で再現していることが示されており、LLM を用いた精神疾患シミュレーションの妥当性が確認されている [7]。

Aster らは、LLM 模擬患者が生成する共感的表現が学習者に与える影響を分析している。医学部学生を対象に、循環器疾患を想定したシナリオに基づき、ChatGPT をバーチャル患者として用いたテキストベースの医療面接を実施した。対話内容は、共感的コミュニケーション符号化システム（Empathic Communication Coding System: ECCS）を用いて分析され、学生が患者の感情表出や困難の語りに対して暗黙的な共感応答を示す場面が確認された。これらの結果から、LLM を用いたバーチャル患者は、従来の模擬患者訓練を補完する手法として、共感的医療面接能力の育成に一定の教育効果を持つ可能性が示唆されている [8]。

以上のとおり診断や支援を目的として、大規模言語モデルやデジタル技術を活用した研究は進められており、模擬患者やバーチャル患者を用いた医療面接訓練、心理的支援の有効性が示されつつある。また、本研究のように特定の精神疾患特性を有する人格として振る舞う Chatbot を模擬患者として構築し、医療従事者が診断に必要な情報収集や信頼関係構築を意識した対話訓練を行うことを主目的とした研究は十分に行われている。

ない。

3 各種精神疾患特性を持つ Chatbot による対話トレーニングシステム

本節では、各種精神疾患特性を持つ Chatbot による対話トレーニングシステムについて説明する。

3.1 模擬患者とは

本節では、模擬患者について説明する。本研究で我々が開発するシステムにおいて再現の対象としている「模擬患者」[9] とは、1964 年に米国の Barrows がその利用を報告した教育手法であり、医学・薬学教育において患者役を演じるために訓練された人々のことを指す。模擬患者には二つの役割がある。一つは、学習者が医療面接や診察の練習を行うための教育・練習における協力者としての役割である。もう一つは、試験などにおいて、すべての受験者に対して同じ応答になるようにあらかじめ固定された問答などが準備された試験協力者としての役割である。一般的にはこれらを総称して模擬患者と呼ぶことが多いが前者を模擬患者といい、後者を標準模擬患者として区別されている。模擬患者では、学習者の理解度や対応に応じて柔軟な演技やフィードバックが求められる一方、標準模擬患者では、事前に定められたシナリオに基づき、一定の反応を再現することが求められる。

模擬患者を用いた学習を通じて、医療従事者は主に三つのスキルを身につけることができる。第一に、患者から話を聞き、必要な情報を整理する力である。これは、患者の訴えや症状を医学的に必要な情報として整理しながら聞き出す力と、患者が安心して話せる雰囲気をつくる態度の両方を身につけることを意味する。第二に、身患者に配慮しながら体を診るための対応力、診察の前に声をかけるなどして患者の不安を和らげ、無理のない形で体勢を誘導するなど、人に配慮した実践的な対応を学ぶ。第三に、患者の立場を意識する力の向上である。診察後に模擬患者から感じた不安や安心感についてフィードバックを受けることで、自分の言葉遣いや態度を振り返り、より良い対人対応を身につけていくことができる。

本研究では模擬患者が担ってきた「患者としての振る舞い」を精神疾患特性に基づいて Chatbot 上に再現することで、安全かつ反復可能な環境において、医療従事者が段階的に対話能力を高めるための訓練基盤の構築を目指す。

3.2 Chatbot による対話トレーニングシステムの概要

本システムは、ユーザ（医療従事者）、疾患特性を付与した Chatbot、および対話インタフェースから構成される。ここで Chatbot は、精神疾患患者を想定した模擬患者として振る舞い、医療従事者に対して対話訓練の対象を提供する役割を担う。ChatBot を用いた提案手法の概要図を図 1 に示す。まず、ユーザは訓練対象とする精神疾患を選択する。次に、選択された疾患に基づき、当該疾患特有の症状や対話傾向を反映した模擬患者 Chatbot が生成される。

ユーザは医師または医療従事者の役割として模擬患者 Chat-

ChatBotとの会話による医療従事者向け対話訓練システム

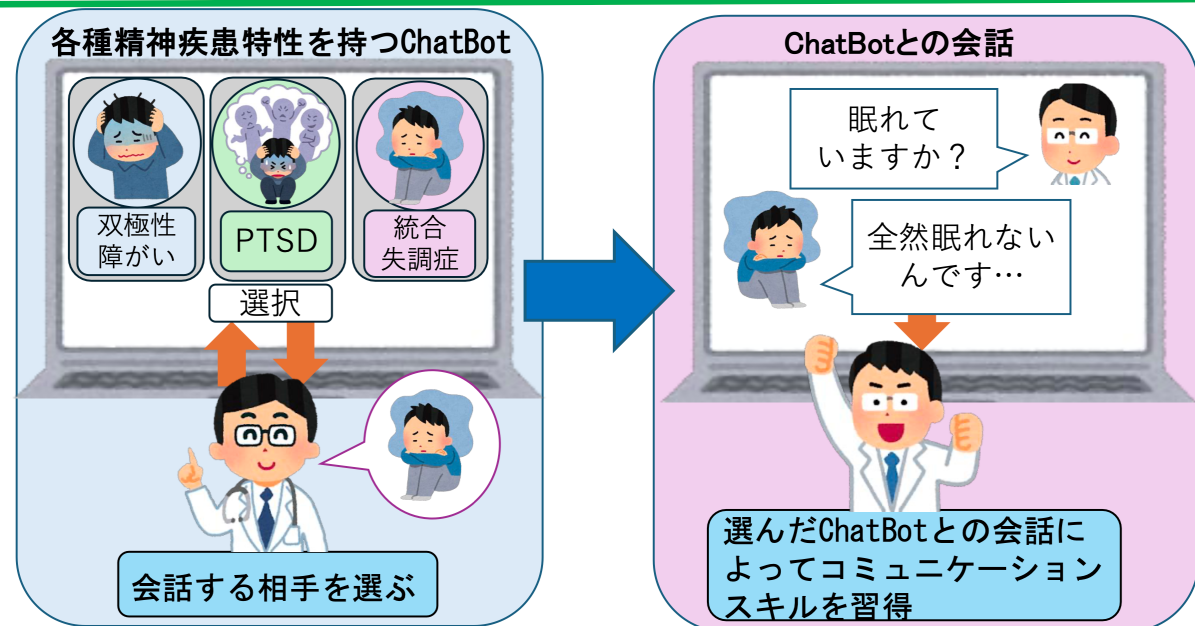


図1 ChatBot との会話による医療従事者向け対話訓練システムの概要図

bot との対話を開始し、実際の患者対応を想定した訓練を行う。これにより、ユーザは模擬患者の発言内容に応じて、症状を引き出すための質問の組み立てや質問順序を工夫し、必要に応じた追加質問や共感的な言語応答を行うことが求められる。このような対話を通じて、実際の診療場面を想定した基本的な問診スキルの習得が可能となる。

3.3 各種精神疾患特性の Chatbot の生成方法

本システムでは、精神疾患患者の模擬患者として振る舞う Chatbot を構築するために、大規模言語モデル (LLM) に対してプロンプトエンジニアリングを用いて疾患特性を付与する。具体的には、うつ病、双極性障害、統合失調症、PTSD、アルコール依存症の5つの精神疾患を対象とした。プロンプトの作成にあたっては、医療学生の教育を目的とした模擬医療診療の YouTube 動画 [10] [11] [12] [13] [14] の内容を参考に、各疾患に見られる臨床的な対話傾向（例：双極性障害における多弁や思考の飛躍、PTSD におけるトラウマへの回避、アルコール依存症における否認など）をペルソナとして設定した。また、年齢や背景情報、主訴などの条件を詳細に記述することで、現実の臨床場面に近い、多様な患者像を持つ Chatbot を生成可能とした。

4 疾患特性の再現性に関する評価実験

本節では、構築した Chatbot が特定の精神疾患特性をどの程度表現できているかを確認するための実験について述べる。

4.1 実験手順

構築した各種精神疾患の Chatbot に対し、プロンプト作成時に参考にした模擬医療診療の YouTube 動画で医師役が行っ

ていたのと同じ診察質問を入力とした。これに対し Chatbot が生成した回答（生成文）と、実際の Youtube 動画内で模擬患者が発話した回答（参照文）を比較することで、疾患特性の再現度を評価した。各種精神疾患の Chatbot に対して行った診察質問回数は実際の模擬医療診療の Youtube 動画で行われた診察質問回数と同じ回数行った。

4.2 評価方法

評価方法は、各種精神疾患を持つ Chatbot の出力（回答）と実際の模擬医療診療の Youtube 動画の模擬患者の回答に対して BERTScore を用いて評価した。BERTScore とは、事前学習済み言語モデル BERT を利用して、生成文と参照文の類似度を算出する評価指標である。単語の表層的な一致を見る n-gram ベースの指標とは異なり、文脈を考慮した埋め込みベクトルを用いることで、単語が異なっても意味的に近い表現であれば高く評価される点が特徴である。なお、BERTScore は文の意味的類似度を分散表現に基づいて評価する指標であり、応答の臨床的妥当性や対話としての自然さを直接評価するものではない。本実験では、模擬患者の回答（正解）の意味内容を Chatbot がどの程度捉えているかを測るため、この指標を採用した。

4.3 実験結果

各精神疾患（うつ病、自閉症、統合失調症、PTSD、アルコール依存症）の Chatbot について、模擬診療動画の模擬患者回答（正解）と Chatbot 生成回答の類似度を BERTScore で算出した。表1に、各疾患における BERTScore の平均値、最大値、最小値を示す。

うつ病に関する全質問の平均 BERTScore は 0.880 であり、全

体の中で最も高い数値を示した。一方で、統合失調症やPTSDなどの疾患では、平均 BERTScore が 0.70 台となり、うつ病と比較してやや低い傾向が見られた。全疾患において平均 0.7 以上を記録しており、一定の意味的類似性が確認された。

表 1 各疾患 Chatbot における BERTScore の評価結果

疾患	平均	最大	最小
うつ病	0.880	0.928	0.838
自閉症	0.731	0.880	0.630
統合失調症	0.706	0.821	0.613
PTSD	0.708	0.811	0.638
アルコール依存症	0.732	0.822	0.667

4.4 考 察

実験の結果、うつ病では平均 0.87 以上と高いスコアを示し、その他の疾患でも平均 0.7 以上となり、構築した Chatbot が模擬患者の発話内容の意味的特徴をおおむね再現できていることが確認できた。一方で、本実験で用いた BERTScore は、生成文と参照文の意味的類似度を分散表現上で算出する指標であり、応答の臨床的妥当性や疾患構造そのものの再現度を直接評価するものではない。例えば、統合失調症において妄想の存在自体は再現されていても、妄想の具体的内容が参照文と異なる場合、疾患特性としては妥当であってもスコアは低下する。また、PTSD における回避的応答のように、症状として自然な情報制限が行われた場合にも、参照文との語彙重なりが減少することで数値は抑制され、つまり、疾患特性として妥当な応答であっても、参照文との意味的類似度が高くなるとは限らない。

4.4.1 Chatbot の応答特性と課題

具体的には、うつ病では、表 2 に示すように、正解文と生成文のいずれにおいても「楽しめることがない」「以前はできていたが今はできない」といった無快感の内容がほぼ同様の構造で表現されており、症状の核心が適切に再現されていることが分かる。PTSD の事例において、「詳しいことは話せませんが」と回答を拒む「回避」の症状が再現されており、アルコール依存症の事例では、「飲みすぎだとは思いません」と自身の飲酒問題を認めない「否認」の防衛機制が明確に表現されていた。また、統合失調症においても、「誰かに狙われている」といった被害的な妄想を含む発言が見られ、現実を正しく判断する力が不安定になっている様子が確認された。これらの結果から、正解データ（動画のシナリオ）と一言一句一致していなくとも、模擬患者としての対話訓練においてリアリティのある反応が生成されていることが確認された。一方で、一部の疾患においては課題も確認された。特に自閉症の事例において、実際の患者が「誰も『応募させた』わけじゃない」と社会的文脈を無視した字義通りの回答をしたのに対し、Chatbot は「仕事が必要だからです」といった社会的妥当性の高い回答を行う傾向が見られた。統合失調症では、妄想という構造自体は維持されているものの、参照動画と同一の妄想テーマや具体的エピソードが対話全体を通して保持されないケースが確認された。

これは、基盤となる LLM が「人間に役立つ・適切な」応答をするように調整されていることに起因すると考えられる。特定の精神障害特有の「社会的コミュニケーションの障害」や「文脈の不一致」を再現するためには、制約条件をプロンプト設計に組み込むことが有効であると考えられる。また、PTSD における回避症状やアルコール依存症における否認傾向のように、情報を制限する応答スタイルを明示的に強調することで、疾患特有の発話パターンをより安定的に生成させることが期待される。

4.4.2 評価指標としての BERTScore の有効性と限界

BERTScore は文の意味的な類似度を測る指標であり、うつ病の事例（表 2）のように、Chatbot が正解データと同様の文脈で症状を訴えた場合には高い値（0.928）を示した。このことから、生成文の流暢さや大まかな意味的整合性を確認する評価指標としては一定の有用性が認められる。しかし、対話訓練システムの評価指標として用いる場合、BERTScore の数値が必ずしもシステムの質を反映しない側面が明らかとなった。その主な要因は、BERTScore が「臨床的な妥当性」ではなく「テキストの意味的な近さ」を算出する点にある。例えば、PTSD やアルコール依存症において、Chatbot が「話したくない（回避）」や「問題ない（否認）」といった短い応答や拒絶反応を示した場合、正解データが事実を詳細に説明する内容であれば、意味内容の乖離により BERTScore は低く算出される（PTSD 最低値: 0.6382 など）。しかし、前述の通りこれらは疾患特性としては正しい振る舞いであり、訓練システムとしては肯定的に評価されるべき挙動である。また、うつ病の最低値（0.8380）の事例に見られるように、正解データに含まれる固有の背景設定（「スーパーの買収」「通信販売の督促」など）を Chatbot が生成できない場合にもスコアが低下するが、これは症状の再現度とは直接関係のないペナルティといえる。以上のことから、精神疾患特有の症状の再現度を測る指標として、BERTScore のような自動評価指標のみでは不十分であることが示された。今後は、自動評価指標の数値向上のみを目指すのではなく、症状リストに基づいたチェックリスト評価する手法や、最終的には専門的知識を持った評価者が対話ログを確認し、適切な応答であるかを評価する方法が考えられる。

表 2 うつ病 Chatbot の BERTScore 最高・最低の回答事例

種別	Score	質問	正解 (模擬患者)	生成 (Chatbot)
最高	0.928	以前は楽しんでいて、今でも楽しめていることはありますか？	本当に何もありません。子供のこともちゃんとできていないし。以前は外出するのが好きで、友達と映画に行ったりしていましたが、今はもう、わざわざそんなことをする気になれません。	いいえ、今はほとんど楽しめることがありません。以前は外出や友達と映画に行くのが好きでしたが、今はその気にもなれません。
最低	0.838	職場ではどのように過ごされていますか？	あまり行けていません。ただだ、疲れ果ててしまっていて……。それに、スーパーが買収されて賃金がカットされて。支払いの問題もあって。通信販売の会社から督促の手紙が届いたり……。供たちは新しいゲームとかを欲しがっているし……。	出勤しても長く働けず、すぐに倦怠感で動けなくなる日が多いです。周囲の視線を気にしてしまい、黙々と作業をこなすのが精一杯で、ミスも増えています。

5 ま と め

本稿では、医療従事者向け対話訓練システムのための各種精神疾患特性を持つ Chatbot の構築を行った。BERTScore を用いた評価の結果、高い類似度スコアが得られ、構築した Chatbot が各精神疾患の特性を反映した応答を生成できることを示した。一方で、実験を通じて自動評価指標の限界も明らかになった。PTSD における「回避」やアルコール依存症における「否認」といった臨床的に妥当な反応が、正解データとのテキスト記述の乖離により、BERTScore 上では低く評価される傾向が確認された。すなわち、対話訓練システムの評価においては、単なるテキストの類似度だけでなく、症状としての正しさを測る視点が不可欠であることが示唆された。今後は、これらの評価結果を基にプロンプト設計を改善し、対話全体で一貫した疾患特性を維持できる Chatbot の構築を目指す。

謝 辞

本研究の一部は、科研費（課題番号：24K02977）および京都産業大学先端科学技術研究所（人間情報学研究センター）共同研究プロジェクト (M2301)、京都産業大学情報理工学部デジタル制作・研究共創チャレンジプログラムの助成を受けたもので

ある。ここに記して謝意を表す。

文 献

[1] 厚生労働省. 精神保健医療福祉施策の現状と課題. <https://www.mhlw.go.jp/content/11121000/001374464.pdf>, 2024. 最終閲覧日：2026 年 1 月 7 日.

[2] Howard S. Barrows. Simulated patients in medical teaching. *Canadian Medical Association Journal*, 98:674–676, 1968.

[3] 藤崎 和彦 and 尾関 俊紀. わが国での模擬患者 (sp) 活動の現状. *医学教育*, 30(2):71–76, 1999.

[4] 本田 芳香 and 塚越 フミエ. 模擬患者導入による学習の有効性. *東京女子医科大学看護学部紀要*, 4:33–38, 2002.

[5] Akira Yamamoto, Masahide Koda, Hiroko Ogawa, Tomoko Miyoshi, Yoshinobu Maeda, Fumio Otsuka, and Hideo Ino. Enhancing medical interview skills through ai-simulated patient interactions: Nonrandomized controlled trial. *JMIR Medical Education*, 10:e58753, 2024.

[6] 遠山 修平, 井上 貴夫, 村井 浩一, 川尻 真也, and 小林 透. 大規模言語モデルを活用した模擬患者の ai モデル化手法の提案. In *2025 年度 人工知能学会全国大会 (第 39 回)*, pages 1Win4–79, 2025.

[7] Jingoo Lee, Kyungho Lim, Young-Chul Jung, and Byung-Hoon Kim. Psyche: A multi-faceted patient simulation framework for evaluation of psychiatric assessment conversational agents. *arXiv preprint arXiv:2501.01594*, 2025.

[8] Alexandra Aster, Sophia Viktoria Ragaller, Tobias Raupach, and Ambra Marx. Chatgpt as a virtual patient: Written empathic expressions during medical history taking. *Medical Science Educator*, 35:1513–1522, 2025.

[9] 中野 重行. 医療コミュニケーションの学習と模擬患者 (sp). *薬理と治療*, 38(12):1077–1087, 2010.

[10] Youtube. 教育のための精神医学面接：うつ病. <https://www.youtube.com/watch?v=4YhpWZCdIZc>, 2012. 閲覧日：2026-01-07.

[11] Youtube. Alcohol dependence syndrome | 2012. <https://www.youtube.com/watch?v=QmFwph3ql3w>, 2012. 閲覧日：2026-01-07.

[12] Youtube. General adult psychiatry mse ptsd single station. <https://www.youtube.com/watch?v=FGK-zisfHEE>, 2012. 閲覧日：2026-01-07.

[13] Youtube. Jan role-playtrainingseries:interview challenges. <https://www.youtube.com/watch?v=VM5EmNX9T2I>, 2017. 閲覧日：2026-01-07.

[14] Youtube. 教育のための精神医学面接：精神病. <https://www.youtube.com/watch?v=ZB28gfSmz1Y>, 2012. 閲覧日：2026-01-07.