

厚生労働科学研究費補助金（がん対策推進総合研究事業）
科学的根拠に基づくがん情報の提供及び均てん化に向けた体制整備に資する研究（23EA1026）
（分担研究報告書）

AI-Chatbotでの情報活用を促進するためのwebがん診療ガイドラインの形式

研究協力者 西迫 宗大 国立がん研究センター がん対策研究所 がん情報提供部（特任研究員）
研究分担者 東 尚弘 東京大学大学院 医学系研究科 公衆衛生学分野（教授）
研究代表者 若尾 文彦 国立がん研究センター がん対策情報センター本部（副本部長）

研究要旨

正確な情報を参照し返答する検索拡張生成 AI（RAG-AI Chatbot）は、医療情報の提供手段として有用とされる。本研究では、Web 上に公開されているがんのガイドラインを参照し返答する AI の開発に有用な Web ガイドラインのサイト構造を検討した。医師向けおよび患者向けのガイドライン 5 編のサイト構造を確認し、各ガイドラインの目次ページを対象に、ページ内の全 URL を取得するスクレイピング処理（Python, BeautifulSoup ライブラリ）を行い、取得した URL のページ内容を確認した。ガイドライン本体が含まれているページとそれ以外のページに分類し、全抽出 URL 数に対する、重複ページ数とガイドライン本体に該当する URL 数の割合を算出し、どのガイドラインが Clinical Question (CQ) レベルで効率的に抽出できるのかを確認した。その結果、「患者さんと家族のための肺がんガイドブック 2024 年版」では、ガイドラインコンテンツに該当する URL の割合が 5 編の中で最も高く（105/106, 99%）、さらに CQ レベルでの抽出が可能であった。一方、外部リンクや章ページのアンカーを通じてガイドラインコンテンツを提示しているサイト構造は、ガイドライン本体の抽出の効率が悪く（0-10%）、さらに CQ 単位での抽出が困難であった。ガイドラインを読み込み返答するシステムにおいて、多くのがん種に応用するためには、複数のガイドラインを横断的に読み込む必要がある。しかし、現在の Web ガイドラインは学会ごとに独自のサイト構造で運用されており、AI による参照を目的とした CQ の抽出には個別の対応が必要である。将来的に Web ガイドラインのサイト構造が標準化されれば、同一のコマンドで複数のガイドラインから同時に情報を取得することが可能となり、RAG-AI Chatbot によるがん情報の横断的な提供が現実的になると考えられる。

A. 研究目的

インターネットやソーシャルメディアの発達により、患者が医療情報を取得し、自らの治療方針を検討する機会は大きく拡大した。一方で、Web 上には根拠の不確かな情報や誤解を招く情報も多く存在し、特にがん医療においては、誤った情報に基づく判断が、適切な治療機会の喪失や経済的負担につながる問題が指摘されている。こうした課題に対し、近年では AI による医療情報の提供が注目され、信頼性の高い情報ソースを参照して回答を生成する検索拡張生成型 AI（Retrieval-Augmented Generation, RAG）の有用性が示されている。Web 版のがん診療ガイドラインは将来的に AI が参照情報として利用されていくと考えられる。

AI による情報抽出の観点では各学会が公開している Web ガイドラインは検討されていない。とくに、Clinical Question (CQ) を含むガイドラインページに対応する情報がページ単位でどのように配置されてい

るかは、AI が情報を効率的に読み取り、精度の高い返答を生成する上で重要であると考えられるが、その標準的な形式は提唱されていない。

そこで本研究では、がん医療ガイドラインを AI が読みやすく構造的に整備するための基礎的検討として、公開されている複数の医師向け・患者向けガイドラインの Web サイト構造を分析し、目次ページからの URL スクレイピングを通じて、CQ 情報が含まれるガイドラインページの抽出効率を評価した。これにより、RAG-AI チャットボットの開発において、より効率的かつ汎用的に活用可能な Web ガイドライン構造の要件を明らかにすることを目的とした。

B. 研究方法

解析の対象

事前に研究許可を取得した 5 つのガイドラインを研究の対象とした（表 1）。乳癌診療ガイドライン 2022 年版 治療編（編者；日本乳癌学会編、目次ペ

ージURL ; <https://jbcs.xsrv.jp/guideline/2022/c/>)、患者さんと家族のための肺がんガイドブック2024年版 (日本肺癌学会 ; <https://www.haigan.gr.jp/public/guidebook/2024/>)、肺癌診療ガイドライン 悪性胸膜中皮腫・胸腺腫瘍含む2024年版 (日本肺癌学会 ; <https://www.haigan.gr.jp/publication/guideline/examination/2024/>)、2024年度版甲状腺腫瘍診療ガイドライン (日本内分泌外科学 ; <http://jaes.umin.jp/guideline/files/guideline2024.pdf>)。また、癌治療学会が運営するがん診療に関連するガイドラインをまとめたサイト「がん診療ガイドライン」より、がんのリハビリテーション診療ガイドライン 第2版 (日本リハビリテーション医学会 ; <http://www.jsco-cpg.jp/rehabilitation/>) を対象とした。

サイト構造の確認

解析対象ガイドラインのホームページのURL、目次ページのURL、Webガイドラインを掲出しているサイトの運営元、表示形式 (HTML、PDF) を記録した (表1)。ガイドラインページにアクセスし、ユーザーインターフェイス (UI) および使用感を記録した後、HTMLソースを確認して、Webガイドラインのディレクトリおよびガイドライン本体コンテンツのファイル位置を確認した。

目次ページからのガイドライン本体の抽出と評価

目次ページを対象にPython (BeautifulSoup ライブラリ、図1.) でスクレイピング処理を行い、ページ内に存在する全URLのリストを取得した。ページにアクセスし内容を確認し、ガイドラインページ/それ以外のコンテンツについて分類した。さらに、同一のURLもしくは異なるURLで同一のコンテンツを指定しているパターンを確認し、これらをまとめて「重複ページ」として記録した。評価は、全抽出URL数に対する、重複ページ数とガイドライン本体のURL数の割合を算出し、どのガイドラインがコンテンツを細かく (CQレベルで) 効率的に抽出できるのかを確認した。

標準的なディレクトリ案

スクレイピングの結果とサイト構造を比較し、最も効率的にガイドラインコンテンツを抽出できるWebサイト構造を検討した。ガイドラインのAIによる利用がされやすいディレクトリを標準案として、検討した結果を図示した。

(倫理面への配慮)

本研究は、個人情報を取り扱うことはない。したがって、個人情報保護上は特に問題は発生しないと考ええる。

C. 研究結果

1. サイトの構造

・患者さんと家族のための肺がんガイドブック2024年版

本ガイドラインはHTML形式で表示されており、目次ページのユーザーインターフェイスは、シンプルな視覚であり、書籍の目次ページと似ている視覚であった。クリックすることにより展開し、階層的な構造となっていた。キーワードによるCQの検索が可能であり、また、ガイドライン以外の情報はほぼ見られない構成となっていた。

Webガイドラインのサイト構造は、目次ファイル下にCQ単位で個別に収納されていた (図2, A)。

目次ページのスクレイピングによって抽出されたURLの総数は106個であり、そのうち重複ページ (重複したURL/コンテンツ) はなく (0%)、内容がガイドラインであったものは105個であった (99%)、ガイドライン以外のページであったURLは1つであり、学会ホームページトップへのリンクであった (図2, B)。CQレベルでのガイドライン本体コンテンツの抽出 (99%) が可能であった。

・肺癌診療ガイドライン 悪性胸膜中皮腫・胸腺腫瘍含む2024年版

本ガイドラインは、HTML形式で表示されていた。目次ページのUIは、シンプルで構成であり、書籍の目次ページと似ている視覚であった。クリックすることにより展開し、階層的な構造となっており、キーワードによるCQの検索が可能であり、ガイドライン以外の情報はほぼ見られない構成となっていた。視覚的には「患者さんと家族のための肺がんガイドブック2024年版」と似ていた。

Webガイドラインのサイト構造は、目次ファイル下に章単位でガイドラインが収納されており、章全体ページをアンカーリンクで個別のCQを指定する構造であった (図3, A)。

目次ページのスクレイピングによって抽出されたURLの総数は384個であり、そのうち重複ページは380個 (90%) であり、それらの内容はすべてがガイドラインであった。ガイドライン以外のページであったURLは4つであり、学会の情報に関するヘッダーリンクであった (図3, B)。ガイドライン本体の抽出は

可能であったが、章をアンカーリンクで指定するWeb構造のために、CQレベルでのガイドライン本体コンテンツの抽出は出来なかった。

・がんのリハビリテーション診療ガイドライン 第2版（癌治療学会「がん診療ガイドライン」）

本ガイドラインはHTML形式で表示されていた。目次ページのユーザーインターフェイスは、ヘッダー・フッターに様々な機能が集約され、さらに、ヘッダーから他のガイドラインにリンクできる機能を認めた。クリックによりガイドラインが展開し操作性に優れ、目次やインターフェイスのデザインがホームページ全体で統一化されており理解しやすい印象であった。

Webガイドラインのサイト構造は、目次ファイルと並列にヘッダーおよびフッターのファイルが存在し、ヘッダーから他がん種へ外部リンクされていた。ガイドラインは章単位で収納されていた（図4, A）。

目次ページのスクレイピングによって抽出されたURLの総数は156個であり、そのうち重複ページは133個（85%）であり、それらの内容はすべてがガイドライン以外であった。ガイドライン本体の取得は15個（9%）であった。ガイドライン以外のページに該当したURLは141個（90%）であった（図4, B）。章単位での収納に加え、ヘッダーに存在する「がん診療ガイドライン」内の他がん種のリストを含み、対象としているがん種以外のがん種の目次ページが抽出されてくる結果となり、目的とするガイドラインのCQレベルでのコンテンツの抽出は出来なかった。

・乳癌診療ガイドライン2022年版 治療編

本ガイドラインはHTML形式で表示されていた。目次ページからガイドライン構造が把握しやすく、構造に沿って操作を誘導するユーザーインターフェイスとなっていた。情報が集約化され、目次ページから様々な情報を取得できるようになっていた。

Webガイドラインのサイト構造は、大項目別（治療編総説・薬物療法・外科療法・放射線療法・略語一覧）にリンクがあり、ガイドライン本体はリンク先のフォルダ別に収納されていた（図5, A）。

目次ページのスクレイピングによって抽出されたURLの総数は71個であり、そのうち重複ページは64個（90%）であり、ガイドライン本体の取得は0個（0%）であった。取得されたURLは学会ホームページ共通のヘッダーやフッターに関連する内容であった（図5, B）。目次ページ内にガイドライン本体の収納がないディレクトリ構造なので、ガイドライン自体の取

得が出来なかった。

・2024年版甲状腺腫瘍診療ガイドライン

本ガイドラインはPDF形式で表示されていた。ユーザーインターフェイスは、電子書籍の閲覧と似ていた。

Webガイドラインのサイト構造は、目次フォルダ内にガイドライン全体（122ページ分）のPDFが収納されていた（図6, A）。

目次ページのスクレイピングでもガイドライン本体PDFの1つのみが抽出され、CQレベルでのガイドライン本体コンテンツの抽出は出来なかった。（図6, B）。

2. 標準的なディレクトリ案

「患者さんと家族のための肺がんガイドブック2024年版」では、CQの含有割合が5編の中で最も高く（105/106、99%）CQの抽出が容易であった（図2.）。本ガイドラインの構造を基に、AI活用を目的としたWebガイドラインの標準的な構造について検討を行った（図7）。CQごとにページが個別に整理されており、さらに共通のヘッダーやフッター、外部リンクを含まないシンプルな構造が、Webガイドラインのテキストデータを最も効率的に取得できるディレクトリ形式として図示された。

D. 考察

本研究では、RAG-AIチャットボットによるがん医療情報の提供を想定し、Web上に公開されている複数のがん診療ガイドラインのWebサイトの構造を解析し、CQを含むガイドラインページのテキストの抽出効率を比較・検討した。その結果、CQごとにページが個別に整理されているWeb構造（「患者さんと家族のための肺がんガイドブック2024年版」）では、スクレイピングにより高い精度でCQを抽出できることが示された（表2.）。一方、章単位のページ構造や外部リンクによる情報提示形式では、CQ単位の抽出が困難であり、AIによる情報活用の妨げとなることが明らかになった。

本研究で使用したスクレイピング手法は、PythonのBeautifulSoupライブラリを用いたごく基本的な処理であった（図1.）。一部のガイドラインでは非常に高い抽出効率を得られたことは、Webガイドラインの構造が適切であれば、AIによる情報取得は低コストかつ高効率で実現可能であることを示している。一方で、現在のWebガイドラインは、学会や発行団体ごとに構造が大きく異なり、AIによる情報活用を前提

とした設計にはなっていない。特に、今回スクレイピングに使用した目次ページのWeb構成に統一性がないため、AIでの処理には個別対応が求められる。このことは情報提供の自動化や汎用化を妨げる要因の1つと考えられた。

各ガイドラインのUIとスクレイピングの結果を比較すると、「人にとって直感的なインターフェイス」と「AIにとって明確な構造」は、それぞれが重視する使いやすさの基準が異なることが分かった。人間の閲覧性を重視した階層的メニューや折りたたみ式UIは、読みやすさには優れるが、AIが要素を構造的に把握し抽出するには適しておらず、逆にAIにとって理想的な構造（CQ単位のHTML分割、共通ヘッダー・フッターの排除、リンクの排除など）は、人間にとっては直感的でない可能性もある。今後、AIを用いたガイドラインの活用を推進するにはユーザーおよびAIの双方にとり利用しやすい構造の提唱がWebガイドライン整備における課題の一つと考えられた。

今後、AIを活用した医療情報提供を社会実装していくためには、Webガイドラインの構造自体を、AIによる活用を前提とした形で標準化する取り組みが重要である。AIが利用しやすいWebガイドラインの標準的な構造が提唱され、すべてのガイドラインが統一された形式で公開されるようになれば、AIによる情報抽出が効率化されるだけでなく、ガイドライン作成時の構造設計にかかる労力の軽減にもつながる。これは、正確な医療情報をすべての人に公平に届けるという医療の本質に根ざした基盤整備であり、医療リテラシーの向上や情報格差の是正にも貢献するものと考えられる。

E. 結論

CQが個別に整理されたWeb構造では高い抽出精度が得られ、章単位や外部リンクによる構造では抽出が困難であった。単純なスクレイピング手法でも、ガイドラインのWeb構造によってCQ抽出効率に顕著な差が生じることから、AIによる医療情報活用にはWebサイト構造の最適化が重要であることが示唆された。将来的にWebガイドラインのサイト構造が標準化されれば、同一のコマンドで複数のガイドラインから同時に情報を取得することが可能となり、RAG-AI Chatbotによるがん情報の横断的な提供が現実的になると考えられる。

F. 健康危険情報

特になし

G. 研究発表

1. 書籍発表 2. 学会発表
なし

H. 知的財産権の出願・登録状況 (予定を含む)

1. 特許取得 なし
2. 実用新案登録 なし
3. その他 なし

資料

表1. 対象としたガイドラインの一覧

ガイドライン名称	目次URL	HPの運営	形式
乳癌診療ガイドライン2022年版 治療編	https://ibcs.xsrv.jp/guideline/2022/c/	学会/団体独自	HTML
患者さんと家族のための肺がんガイドブック2024年版	https://www.haigan.gr.jp/public/guidebook/2024/		
肺癌診療ガイドライン - 悪性胸膜中皮腫・胸腺腫瘍含む2024年版	https://www.haigan.gr.jp/publication/guideline/examination/2024/		
がんのリハビリテーション診療ガイドライン (第2版)	http://www.jscocpg.jp/rehabilitation/	一括 (癌治療学会)	
2024年度版甲状腺腫瘍診療ガイドライン	http://jaes.umin.jp/guideline/files/guideline2024.pdf	学会/団体独自	PDF

```
import requests
from bs4 import BeautifulSoup
from urllib.parse import urljoin

def extract_urls(base_urls):

    all_urls = []
    for base_url in base_urls:
        response = requests.get(base_url)
        response.raise_for_status()
        soup = BeautifulSoup(response.content, "html.parser")
        for link in soup.find_all("a", href=True):
            url = link["href"]

            absolute_url = urljoin(base_url, url)
            all_urls.append(absolute_url)
    return all_urls

# 実行例
base_urls = ["https://XXXXX"]
urls = extract_urls(base_urls)

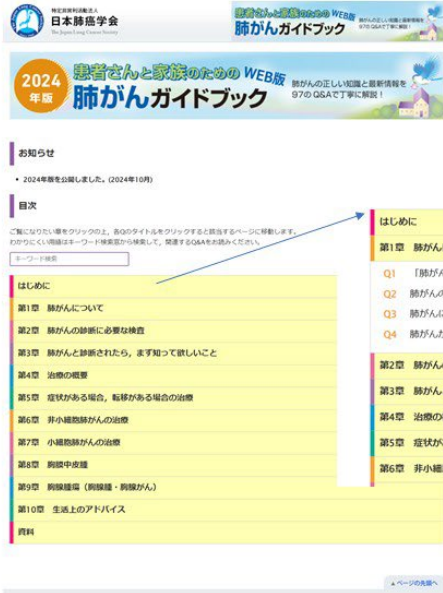
# 取得したURLの一覧を表示
print("取得したURLの一覧:")
for url in urls:
    print(url)
```

図1. スクレイピングに使用したコード (python; BeautifulSoup library)

(A)

患者さんと家族のための肺がんガイドブック2024年版

<https://www.haigan.gr.jp/public/guidebook/2024/>



- ・シンプルであり、実際の目次ページと似ている
- ・クリックで展開し、階層的な構造
- ・順を追って知りたい事を検索
- ・キーワード検索可能
- ・ガイドライン以外の情報はほぼない

(B)

患者さんと家族のための肺がんガイドブック2024年版

<https://www.haigan.gr.jp/public/guidebook/2024/>

抽出されたURL	重複ページ	ガイドライン本体	ガイドライン以外のページ
106	0	105	1

- ・重複ページが無い、ガイドライン本体をほぼ正確に取得した
目次ページはCQの羅列であり、クリックで展開する仕組み
- ・ガイドライン以外のページ取得は学会のHPのトップ
ヘッダーに学会HPのリンクあり

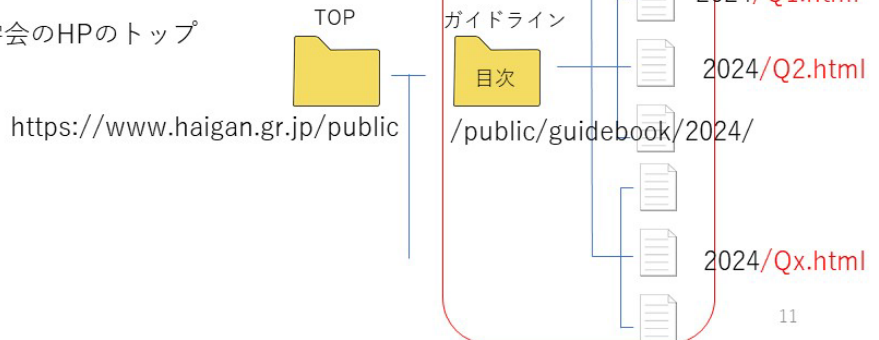


図2. 患者さんと家族のための肺がんガイドブック2024年版 (A) Webガイドライン目次ページのインターフェイス. (B) ガイドライン本体の取得とサイト構造

(A)

肺癌診療ガイドライン-悪性胸膜中皮腫・胸腺腫瘍含む2024年版
<https://www.haigan.gr.jp/publication/guideline/examination/2024/>



- ・シンプルであり、実際の目次ページと似ている
- ・クリックで展開し、階層的な構造
- ・順を追って知りたい事を検索
- ・キーワード検索可能
- ・ガイドライン以外の情報はほぼない



12

(B)

肺癌診療ガイドライン-悪性胸膜中皮腫・胸腺腫瘍含む2024年版
<https://www.haigan.gr.jp/publication/guideline/examination/2024/>

抽出されたURL	重複ページ	ガイドライン本体	ガイドライン以外のページ
384	380	380	4

- ・ガイドライン本体の取得はしているが、重複ページ多かった
CQの表示は、同一の章のページ内でのリンクとなっている
URL自体の重複は存在しなかった
- ・ガイドライン以外のページ取得は学会の情報
ヘッダーに学会HPのリンクあり

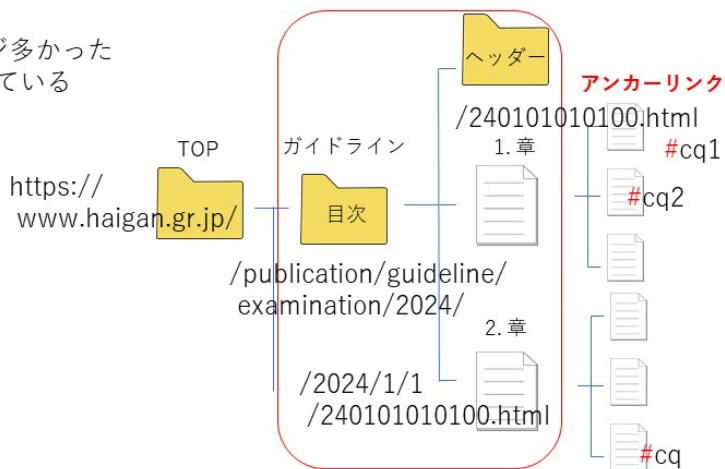


図3. 肺がん診療ガイドライン-悪性胸膜中皮腫・胸腺腫瘍を含む2024年版 (A) Webガイドライン目次ページのインターフェイス. (B) ガイドライン本体の取得とサイト構造

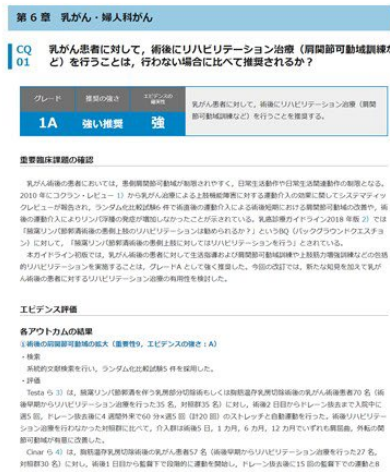
(A)

がんのリハビリテーション診療ガイドライン (第2版)

<http://www.jasco-cpg.jp/rehabilitation/>



- ・ヘッダー・フッターに様々な機能が集約
- ・ヘッダーから他のガイドラインにリンクできる
- ・目次やインターフェイスが統一化されており理解しやすい
- ・クリック展開で操作性がよい



14

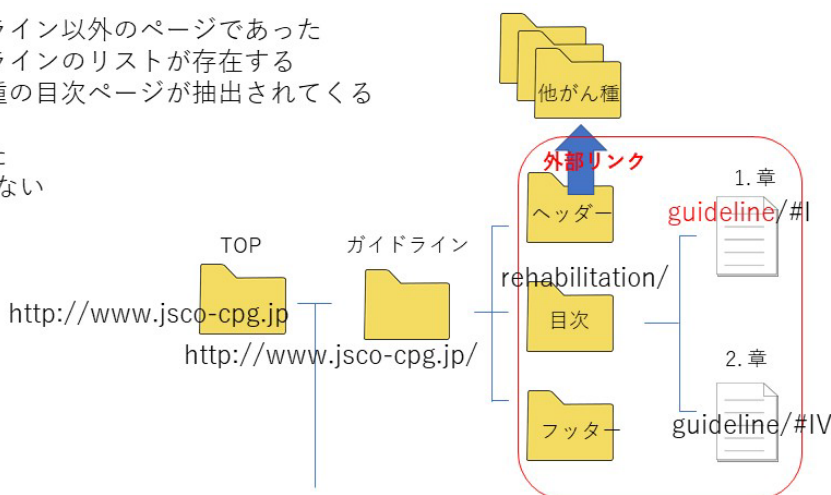
(B)

がんのリハビリテーション診療ガイドライン (第2版)

<http://www.jasco-cpg.jp/rehabilitation/>

抽出されたURL	重複ページ	ガイドライン本体	ガイドライン以外のページ
156	133	15	141

- ・重複ページ多く、ほとんどがガイドライン以外のページであった
ヘッダーにすべてがん種のガイドラインのリストが存在する
対象としているがん種以外のがん種の目次ページが抽出されてくる
- ・ガイドライン本体の抽出が少なかった
章ごとの掲示でCQ毎にはなっていない



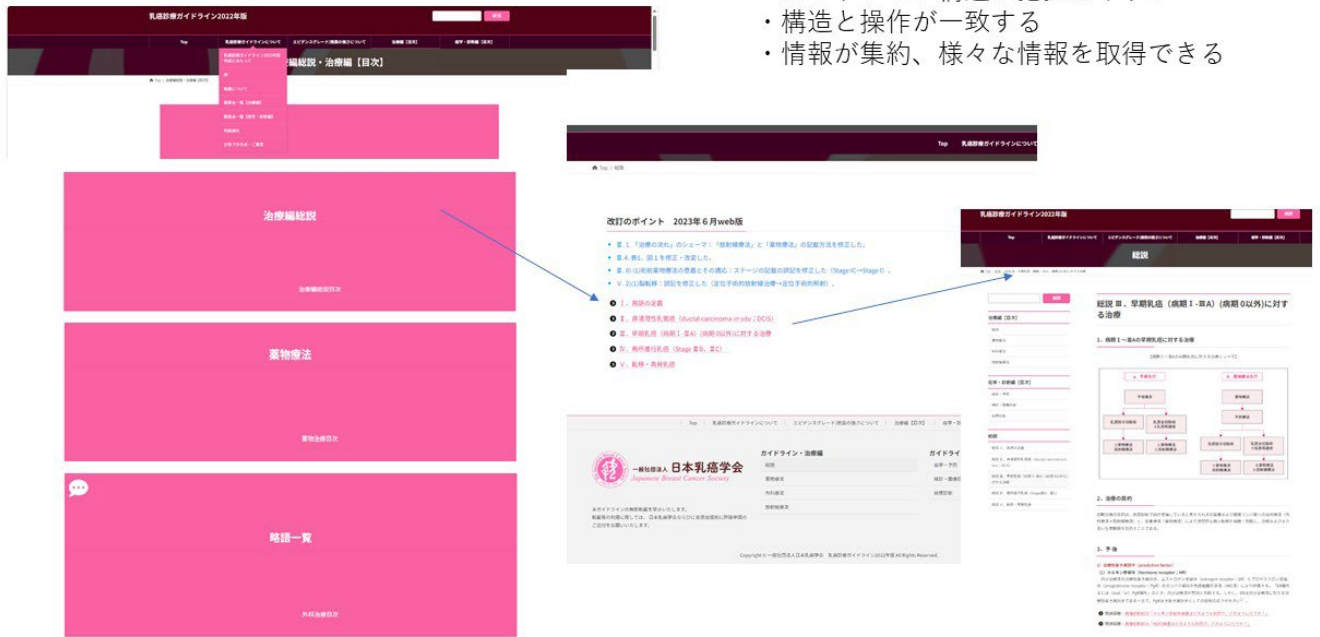
15

図4. がんのリハビリテーション診療ガイドライン (A) Webガイドライン目次ページのインターフェイス. (B) ガイドライン本体の取得とサイト構造

(A)

乳癌診療ガイドライン2022年版 治療編

<https://jbcx.srv.jp/guideline/2022/c/>



- ・ガイドライン構造が把握しやすい
- ・構造と操作が一致する
- ・情報が集約、様々な情報を取得できる

(B)

乳癌診療ガイドライン2022年版 治療編

<https://jbcx.srv.jp/guideline/2022/c/>

抽出されたURL	重複ページ	ガイドライン本体	ガイドライン以外のページ
71	64	0	71

- ・目次ページではガイドライン本体の取得ができなかったリンクやフォルダの構造；
大項目別（治療編総説・薬物療法・外科療法・放射線療法・略語一覧）にリンクがあり、リンク先のフォルダ別に収納

- ・目次ページにガイドライン以外のリンクが多く存在する
学会HP共通のヘッダーやフッターや重複する内容

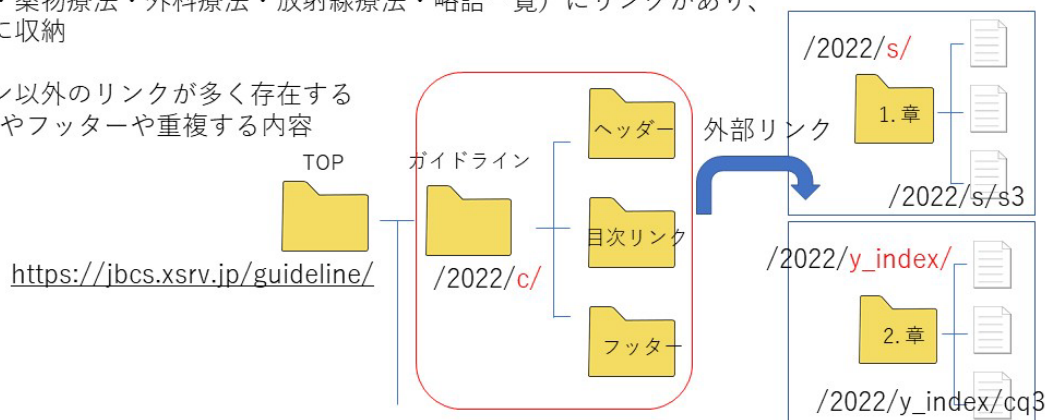


図5. 乳癌診療ガイドライン2022年版 治療編 (A) Webガイドライン目次ページのインターフェイス. (B) ガイドライン本体の取得とサイト構造

(A)

2024年度版甲状腺腫瘍診療ガイドライン

<http://jaes.umin.jp/guideline/files/guideline2024.pdf>



16

(B)

4. 2024年度版甲状腺腫瘍診療ガイドライン

<http://jaes.umin.jp/guideline/files/guideline2024.pdf>

抽出されたURL	重複ページ	ガイドライン本体	ガイドライン以外のページ
1	0	1	0

・ガイドライン全体（122ページ分）が抽出された

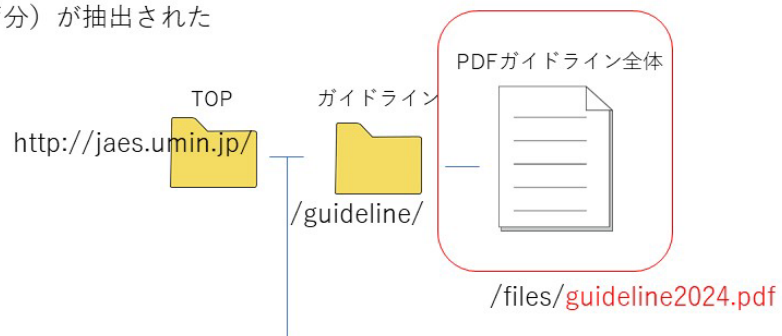


図6. 2024年版甲状腺腫瘍ガイドライン (A) Webガイドライン目次ページのインターフェイス. (B) ガイドライン本体の取得とサイト構造

標準的なディレクトリ構造案

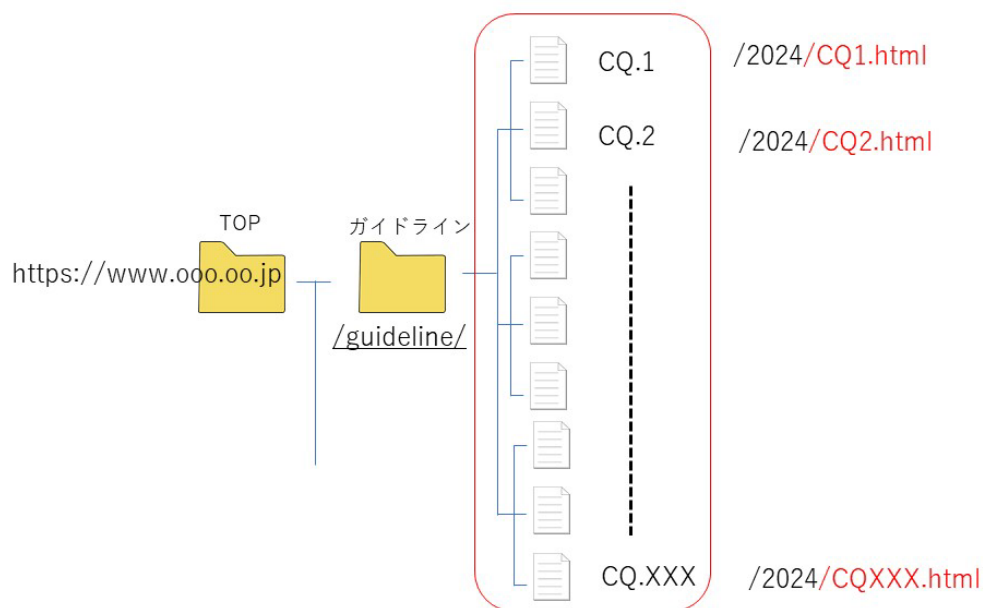


図7. AI活用を目的としたWebガイドラインの構造案