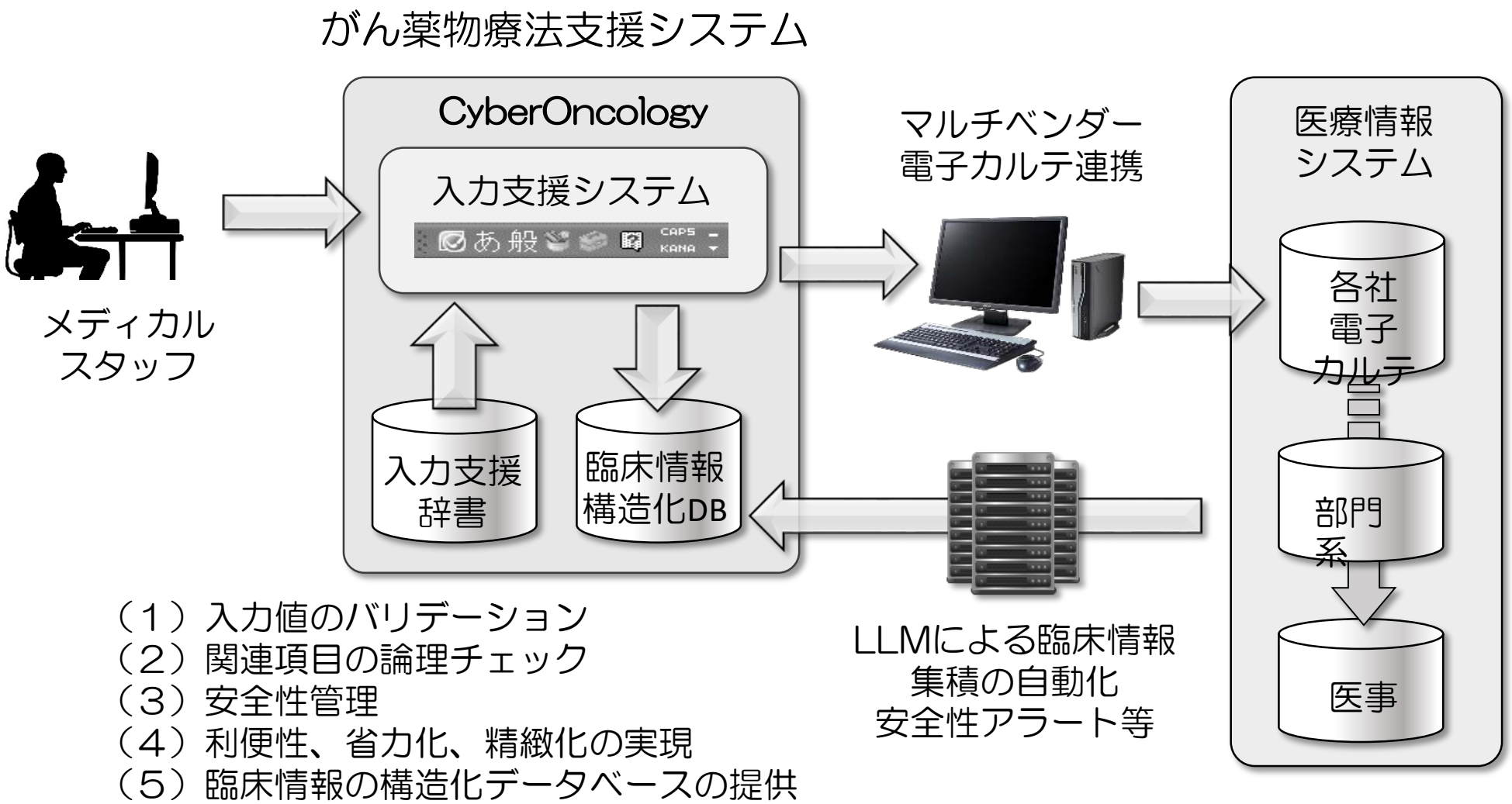
千年カルテデータで鍛える構造化用LLM

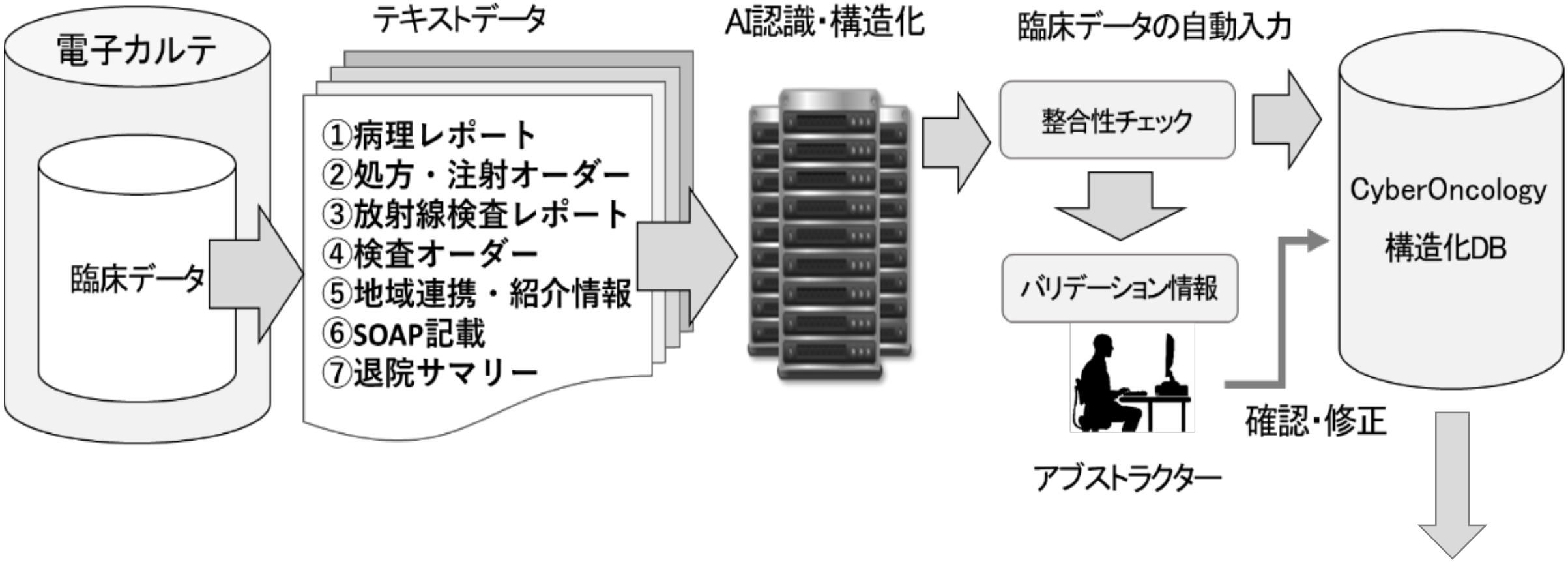
加藤康之[1)], 横田理央[2], 吉原博幸[3)]

1) 新医療リアルワールドデータ研究機構株式会社 (PRiME-R　シニアフェロー)
2) 東京工業大学 学術国際情報センター　教授
3) 京都大学大学院医学研究科 社会健康医学系専攻 健康情報学分野 (京都大学名誉教授)

Cyber Oncologyによるリアルワールドデータ収集環境



電子カルテからの自動抽出のイメージ

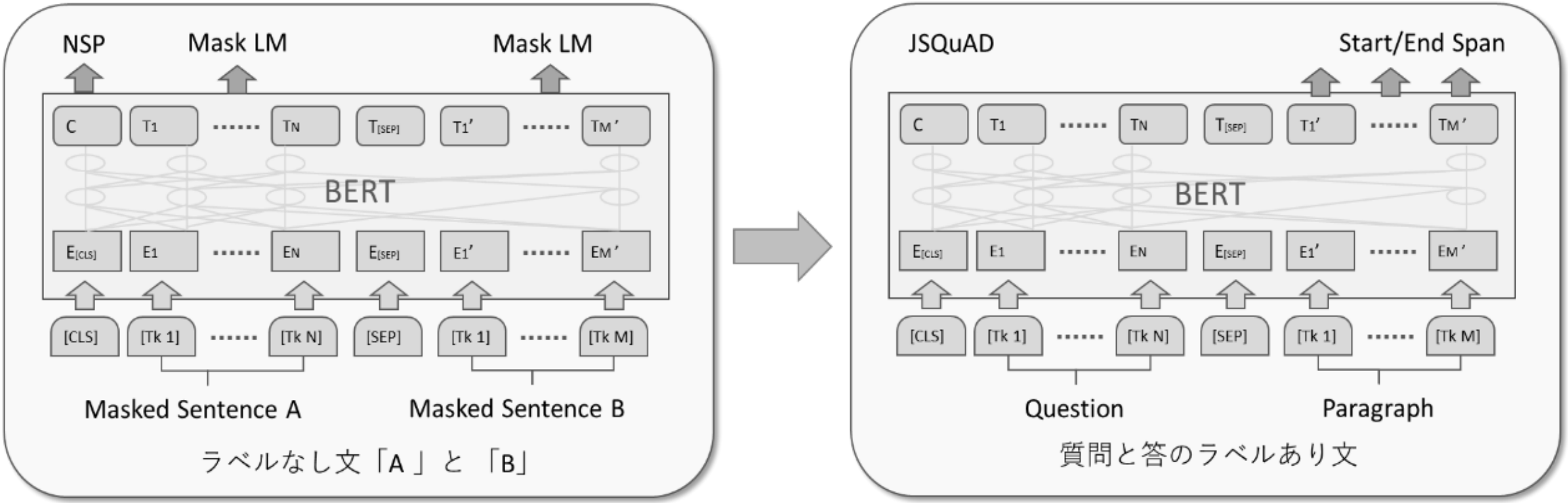


- C-CAT
- 退院サマリー、紹介状（SIP3研究）
- 各学会レジストリ（認定医、専門医、指導医）

当初の検討：BERTモデルの利用（2023/3～月）

PRiME-R

論文投稿：「Cyber OncologyにおけるAI活用の取組みと今後の展望」
Precision Medicine Vol.6 No.7, 2023



事前学習　**（パラメータサイズ：0.2B）**　ファインチューニング

課題：文字数＜256文字、事象の前後把握が難しい　⇒　次ページ

出展：Kyoto University. A BERT published by the Kyoto University.　[cited 2021 May 31]. Available from: http://nlp.ist.i.kyoto-u.ac.jp/EN/.

臨床情報を自動抽出する具体的なイメージ

PRiME-R

現在人間がカルテを読解して臨床情報を抽出しているが
将来、以下のようなイメージでLLMにより自動抽出したい

BERTモデルの課題：文字数＜256文字、事象の前後把握が難しい

カルテ経過記録例

2019年12月初旬に胸痛、背部痛を自覚。近医を受信し左肺門部主流を指摘され、1月7日に当院を紹介受診。
同日に胸腔穿刺を施行し、得られた胸水から肺腺癌と診断
T4N1M1c StageⅣC,BRA,PUL,LYM,PLE,OSS, EGFR(L858R+),KRAS-,BRAF(V600E)-,ROS1-,PDL-1<0%,ALK-
1/31～：C1-7頸椎転移に対し緩和的放射線治療(30Gy/10fr)施行
2/6～：1stline Osimertinib 80mg/dayを開始（28日分）。
11/4：CTにて左肺門部腫瘤、左肺尖結節、肝転移、骨転移増大を認めPDと判断。
11/19～2/10：2nd line CBDCA/PTX/Bev/Atezoを4course施行
2021/3/2：効果判定にて原疾患の増悪を認めPDと判断。
3/16～4/13：3rd line DOC+RAM 2course施行
5/18：効果判定にて両肺の小結節は増加・増大と両測胸水の増加がありPDと判断
5/20：胸水コントロール目的に入院。
5/21：左癌性に対して胸水左胸腔ドレーン挿入
5/24：左胸膜癒着術(ユニタルク4g)を施行。
5/27～：4th line EGFR-TKI rechallenge(Afatinib 20mg/day)開始
5/28：食道の通過障害あり原疾患からの粘膜下への腫瘍浸潤の疑い、まずはAfatinibによる抗腫瘍効果を期待する方針。
・・・【以下数年分続きます】・・・

臨床情報抽出例

項番	臨床項目	小項目	内容（値）
1	病理診断名		肺腺癌
2	病理診断日		不明：病理レポート要確認
3	検体		胸水
4	検体採取日		2020/1/7
5	TNM分類		T4N1M1c
6	Stage		IVC（IVBの間違い）
7	転移部位		BRA,PUL,LYM,PLE,OSS
8	バイオマーカー	EGFR	L858R陽性
9		KRAS	陰性
10		BRAF(V600E)	陰性
11		ROS1	陰性
12		PDL-1	<0%
13		ALK	陰性
14	放射線治療	治療目的	緩和
15		治療部位	C1-7頸椎
16		照射総量	30Gy/10fr
17		治療開始日	2020/1/31
18		治療終了日	不明：放射線治療レポート要確認
19	薬物1次治療	Osimertinib	80mg/day
20		治療開始日	2020/2/6
21		治療終了日	不明：処方オーダー要確認
22	進行増悪日	判断手段	CT
23		増悪日	2020/11/4
24	薬物2次治療	CBDCA/PTX/Bev/Atezo	4course
25		治療開始日	2020/11/19
26		治療終了日	2021/2/10
27	進行増悪日	判断手段	CT
28		増悪日	2021/3/2
29	・・・以下200項目程度		

LLMによる電子カルテ情報の構造化例

PRiME-R

【構造化評価文：電子カルテの経過記録例】

2019年12月初旬に胸痛、背部痛を自覚。近医を受信し左肺門部腫瘤を指摘され、1月7日に当院を紹介受診。
同日に胸腔穿刺を施行し、得られた胸水から肺腺癌と診断
T4N1M1c, StageⅣB, BRA, PUL, LYM, PLE, OSS, EGFR(L858R+), KRAS−, BRAF(V600E)−, ROS1−, PDL−1 < 0%, ALK−
1/31～：C1−7頸椎転移に対し緩和的放射線治療(30Gy/10fr)施行
2/6～：1stline Osimertinib 80mg/dayを開始（28日分）。
11/4：CTにて左肺門部腫瘤、左肺尖結節、肝転移、骨転移増大を認めPDと判断。
11/19～2/10：2nd line CBDCA/PTX/Bev/Atezoを4course施行
2021/3/2：効果判定にて原疾患の増悪を認めPDと判断。
3/16～4/13：3rd line DOC+RAM 2course施行
5/18：効果判定にて両肺の小結節は増加・増大と両測胸水の増加がありPDと判断
5/20：胸水コントロール目的に入院。
5/21：左癌性に対して胸水左胸腔ドレーン挿入
5/24：左胸膜癒着術(ユニタルク4g)を施行。
5/27～：4th line EGFR−TKI rechallenge(Afatinib 20mg/day)開始

LLM

【LLMによる構造化例：Microsoft BingAIの例】


```
[
. . . . . .
  {
    "date": "2020年1月7日",
    "event": "胸腔穿刺を施行し、得られた胸水から肺腺癌と診断",
    "diagnosis": "肺腺癌",
    "treatment": "胸腔穿刺",
    "TNM_classification": "T4N1M1c",
    "stage": "StageⅣB",
    "metastasis": "BRA, PUL, LYM, PLE, OSS",
    "specimen": "胸水",
    "gene_mutation": "EGFR(L858R+), KRAS−, BRAF(V600E)−, ROS1−, PDL−1<0%, ALK−"
  },
  {
    "date": "2020年1月31日～",
    "event": "C1−7頸椎転移に対し緩和的放射線治療(30Gy/10fr)施行",
    "diagnosis": null,
    "treatment": "緩和的放射線治療",
    "TNM_classification": null,
    "stage": null,
    "metastasis": "C1−7頸椎転移",
    "specimen": null,
    "gene_mutation": null
  },
], . . . . . .
}
```


Llama2: 文字数＜4096文字

臨床データ構造化における連関情報の生成とバリデーションのためのデータ管理方式

PRiME-R

患者情報テーブル：T0
- 患者ID
- 生年月日
- 性別
- 血液型、等

1対多

エピソード管理テーブル：T1
- エピソードID
- 患者ID
- がんカテゴリー
- がん情報
- がん部位コード
- 作成日
- 更新日、等

独立がん種毎に作成

1対多

治療ラインテーブル：T2
- 治療ラインID
- エピソードID
- 患者ID
- レジメン名
- 開始日
- 終了日
- 作成日
- 更新日、等

治療ライン毎に作成

1対多

効果判定テーブル：T3
- 効果判定ID
- 治療ラインID
- 効果判定結果
- 判定日
- 文書種別ID
- カルテ記載日
- カルテ記載者
- 検索テキスト

1対多

同意情報テーブル：T5
- 患者ID
- 臨床研究ID
- 同意ステータス
- 同意日

電子カルテ内の記載場所を特定する情報

【文書種別ID】
病理:0，経過記録:1，放射線リポート:2，
注射:3，処方:4，臨床検査:5，サマリー:6，
スキャン:7
【カルテ記載者】医師名等
【検索テキスト】LLM出力の”event”情報等

有害事象テーブル：T4
- 有害事象ID
- 治療ラインID
- 有害事象(CTCAEv5.0)
- グレード
- 発現日
- 終息日
- 文書種別ID
- カルテ記載日
- カルテ記載者
- 検索テキスト

構造化データ抽出機能の評価方式

PRiME-R

事象項番	経過記録例	臨床項目	項目数
1	2019年12月初旬に胸痛、背部痛を自覚。近医を受信し左肺門部腫瘤を指摘され	日時、症状	2
2	1月7日に当院を紹介受診。同日に胸腔穿刺を施行し、得られた胸水から肺腺癌と診断、T4N1M1c StageⅣB, BRA, PUL, LYM, PLE, OSS, EGFR(L858R+), KRAS−, BRAF(V600E)−, ROS1−, PDL−1<0%, ALK−	日時、検体、診断、TNM、Stage、転移(5つ)、バイオマーカー(6つ)	16
3	1/31〜:C1−7頸椎転移に対し緩和的放射線治療(30Gy/10fr)施行	日時、処置、部位、線量	4
4	2/6〜:1stline Osimertinib 80mg/dayを開始(28日分)。	開始日、終了日、治療ライン、治療内容	4
5	11/4:CTにて左肺門部腫瘤、左肺尖結節、肝転移、骨転移増大を認めPDと判断。	日時、効果判定、部位	3
6	11/19〜2/10:2nd line CBDCA/PTX/Bev/Atezoを4course施行	開始日、終了日、治療ライン、治療内容	4
7	2021/3/2:効果判定にて原疾患の増悪を認めPDと判断。	日時、効果判定、部位	3
8	3/16〜4/13:3rd line DOC+RAM 2course施行	開始日、終了日、治療ライン、治療内容	4
9	5/18:効果判定にて両肺の小結節は増加・増大と両測胸水の増加がありPDと判断	日時、効果判定、部位	3
10	5/21:左癌性に対して胸水左胸腔ドレーン挿入	日時、治療内容	2
11	5/24:左胸膜癒着術(ユニタルク4g)を施行	日時、治療内容	2
12	5/27〜:4th line EGFR−TKI rechallenge(Afatinib 20mg/day)開始	開始日、治療ライン、治療内容	3
合計			50

種々のLLMにおける臨床データ抽出機能の評価結果1

PRiME-R

事象項番	項目数	クラウドサービス			オープンソース・英語版			オープンソース・日本語版			
		—	—	—	商用利用可			商用利用不可	商用利用可		
		CharGPT 3.5	Google-BARD	Microsoft BingAI	Llama2 -7B	Llama2 -13B	Llama2 -70B	Weblab -10B	Plamo -13B	llm-jp -13B	Swallow -13B
1	2	2	2	2	2	2	1	0	2	0	2
2	16	16	10	16	0	2	16	8	3	8	12
3	4	4	3	4	0	3	2	0	0	2	3
4	4	3	3	3	2	3	4	0	0	2	3
5	3	3	3	3	0	0	2	0	0	2	2
6	4	3	3	4	2	1	3	0	0	0	3
7	3	3	3	3	0	0	0	0	0	0	2
8	4	3	3	4	2	0	4	0	0	0	3
9	3	0	3	3	0	2	0	0	0	0	2
10	2	2	2	2	0	1	0	0	0	0	2
11	2	2	2	2	0	1	0	0	0	0	2
12	3	3	3	3	2	2	3	0	0	0	3
正解項目数の合計	50	44	40	49	10	15	35	8	5	14	39
構造化率(%)	100	88	80	98	20	30	70	16	10	28	78

種々のLLMにおける臨床データ抽出機能の評価結果2

PRiME-R

Swallowは分担研究者の横田研究室が開発

項番	項目数		オープンソース・英語版			日本純正	Llama2の日本語継続学習					
		BingAI	Llama2-7B	Llama2-13B	Llama2-70B	llm-jp-13B	youri-7B	ELYZA-japanese-Llama-2-7B	ELYZA-japanese-Llama-2-13B	Swallow-13B	japanese-stablelm-instruct-beta-70B	Swallow-70B
1	2	2	2	2	1	0	2	2	0	2	2	2
2	16	16	0	2	16	8	0	5	9	12	9	16
3	4	4	0	3	2	2	0	0	2	3	2	4
4	4	3	2	3	4	2	0	0	3	3	3	3
5	3	3	0	0	2	2	0	0	3	2	3	3
6	4	4	2	1	3	0	0	0	2	3	4	3
7	3	3	0	0	0	0	0	0	3	2	3	3
8	4	4	2	0	4	0	0	0	3	3	4	3
9	3	3	0	2	0	0	0	0	3	2	3	3
10	2	2	0	1	0	0	0	0	2	2	2	2
11	2	2	0	1	0	0	0	0	2	2	2	2
12	3	3	2	2	3	0	0	0	3	3	3	3
正解項目数	50	49	10	15	35	14	2	7	35	39	40	47
構造化率(%)	100	98	20	30	70	28	4	14	70	78	80	94

R6年度厚生労働科学研究費
臨床研究等 ICT 基盤構築・人工知能実装研究事業

大規模言語モデル（LLM：Large Language Model）を活用した医薬品等の有効性・安全性評価のためのアウトカム抽出の方法論の確立に向けた研究（24AC1004）

代表研究者：武藤 学（京都大学・医学研究科　腫瘍内科学講座・教授）

研究の全体像

【高性能LLM選定】 オープンソースの英語版LLMからベースとなるモデルを選定

【日本語対応】 大規模日本語コーパスで継続事前学習

【医療領域対応】 各診療ガイドライン、各取扱い規約、添付文書、最適使用推進ガイドライン、 保険留意事項通知等で継続事前学習

【カルテ自由記載に対応】 大規模電子カルテデータで継続事前学習 (>1TB)

【性能評価】 実臨床データでLLMの構造化精度を評価

ほぼ全ての施設のカルテ記載を理解できるLLMの実現

研究の範囲

EHR

カルテの自由記載文
臨床検査データ
処方データなど

構造化専用 LLM-1

同じ ハードウェア

解析専用 LLM-2

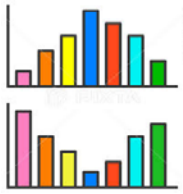
有効性・安全性評価情報等、構造化データ蓄積

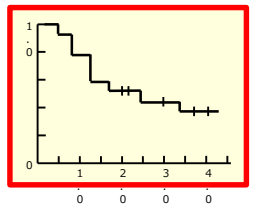
RDB
(～300項目／症例)

本研究成果の目指すところ

自然言語での問い合わせ

解析結果



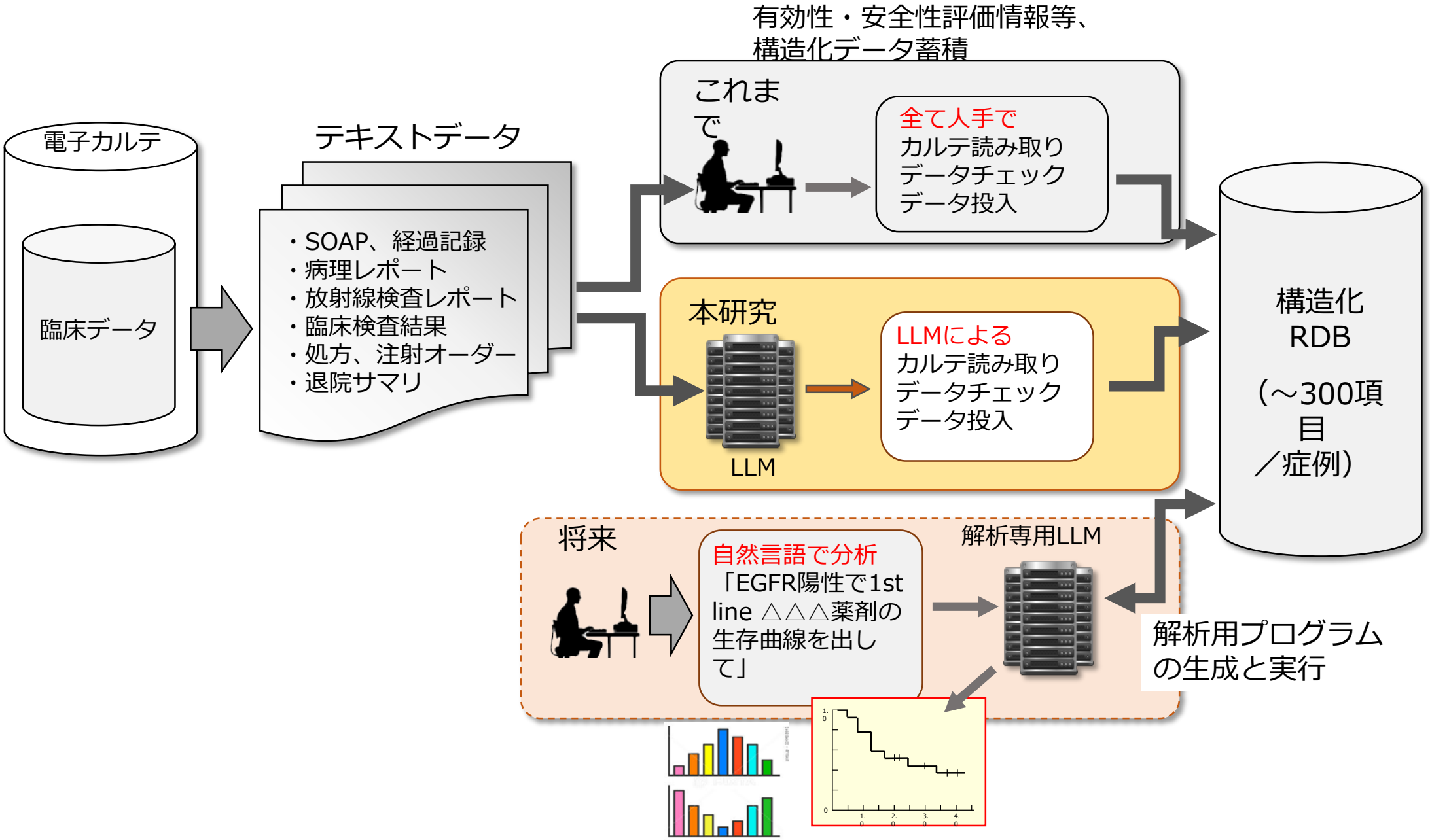


医薬品等の有効性・安全性評価に資するアウトカム抽出を可能とする

解析用プログラムの生成と実行

各種研究レジストリと連携

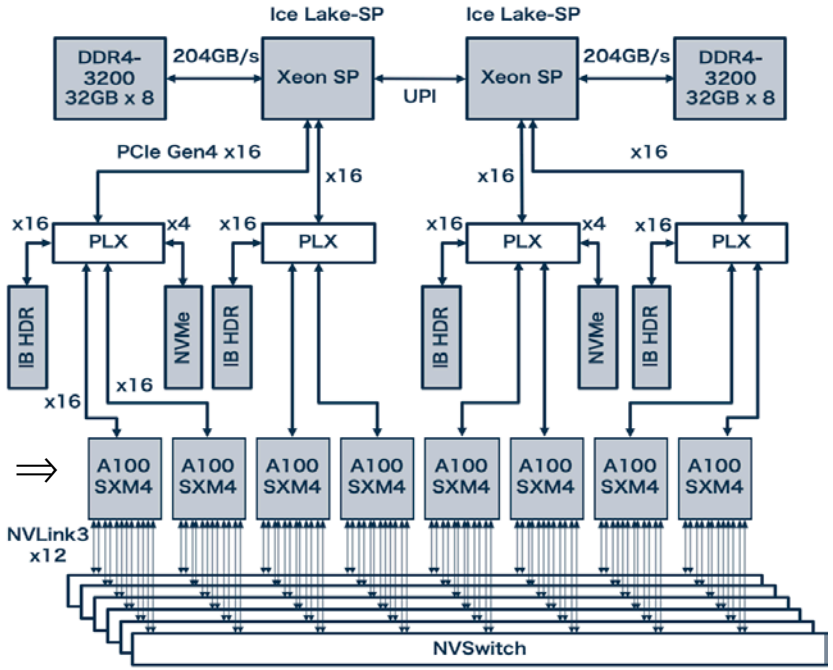
LLMによる臨床データ自動取得のイメージ



産総研ABCIクラスターサーバでの学習

1ノード構成 x 960基

NVIDIA GPU ⇒



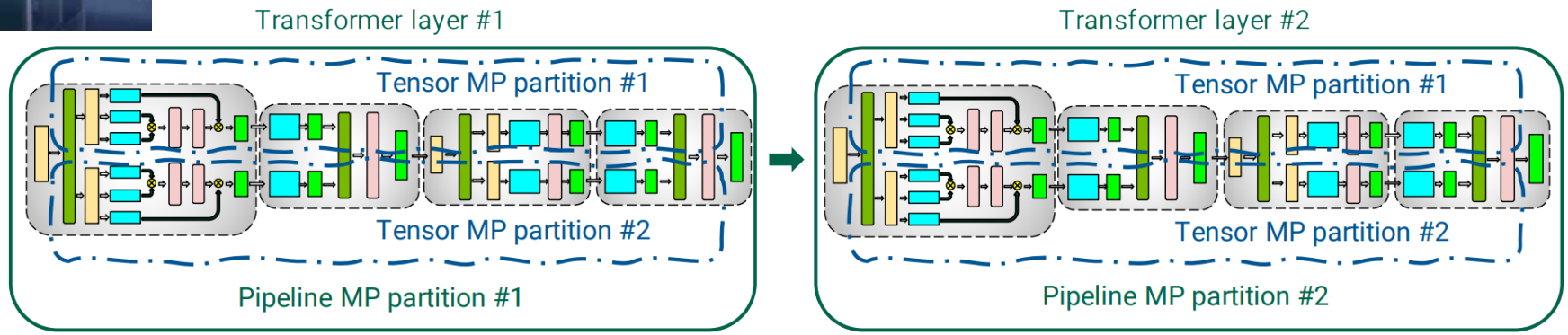


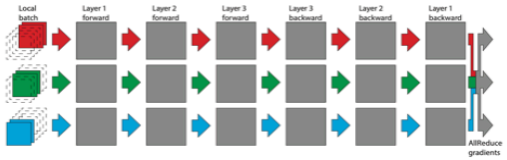
Figure 2: Combination of tensor and pipeline model parallelism (MP) used in this work for transformer-based models.

大規模パラメータLLMの分散学習技術

PRiME-R

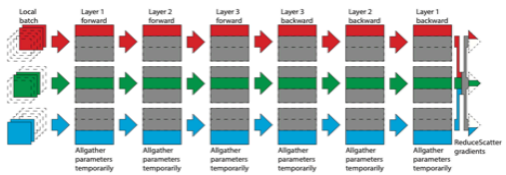
- Llama2-70Bの規模になると4種類の並列化を併用する必要がある
- →これは容易でないため、国内でこの規模のものはSwallowのみ

データ並列 (DP)



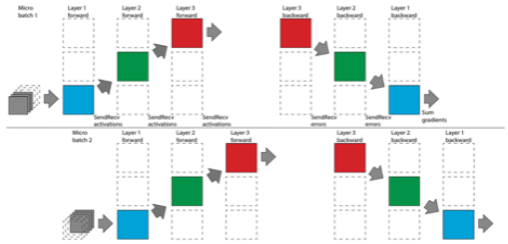
データ：分散
モデル：冗長
通信内容：勾配
通信形式：AllReduce
通信頻度：ステップ毎
長所：実装が簡単
短所：ラージバッチ問題
　　メモリ消費量

ZeRO (FSDP)



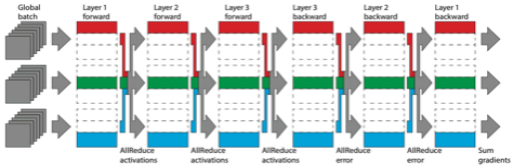
データ：分散
モデル：一時的に分散
通信内容：勾配＋重み
通信形式：ReduceScatter
　　＋AllGather
通信頻度：層毎
長所：実装が簡単
　　省メモリ
短所：ラージバッチ問題

パイプライン並列 (PP)



データ：冗長
モデル：分散
通信内容：活性
通信形式：SendRecv
通信頻度：層毎
長所：省メモリ
短所：パイプラインバブル

テンソル並列 (TP)

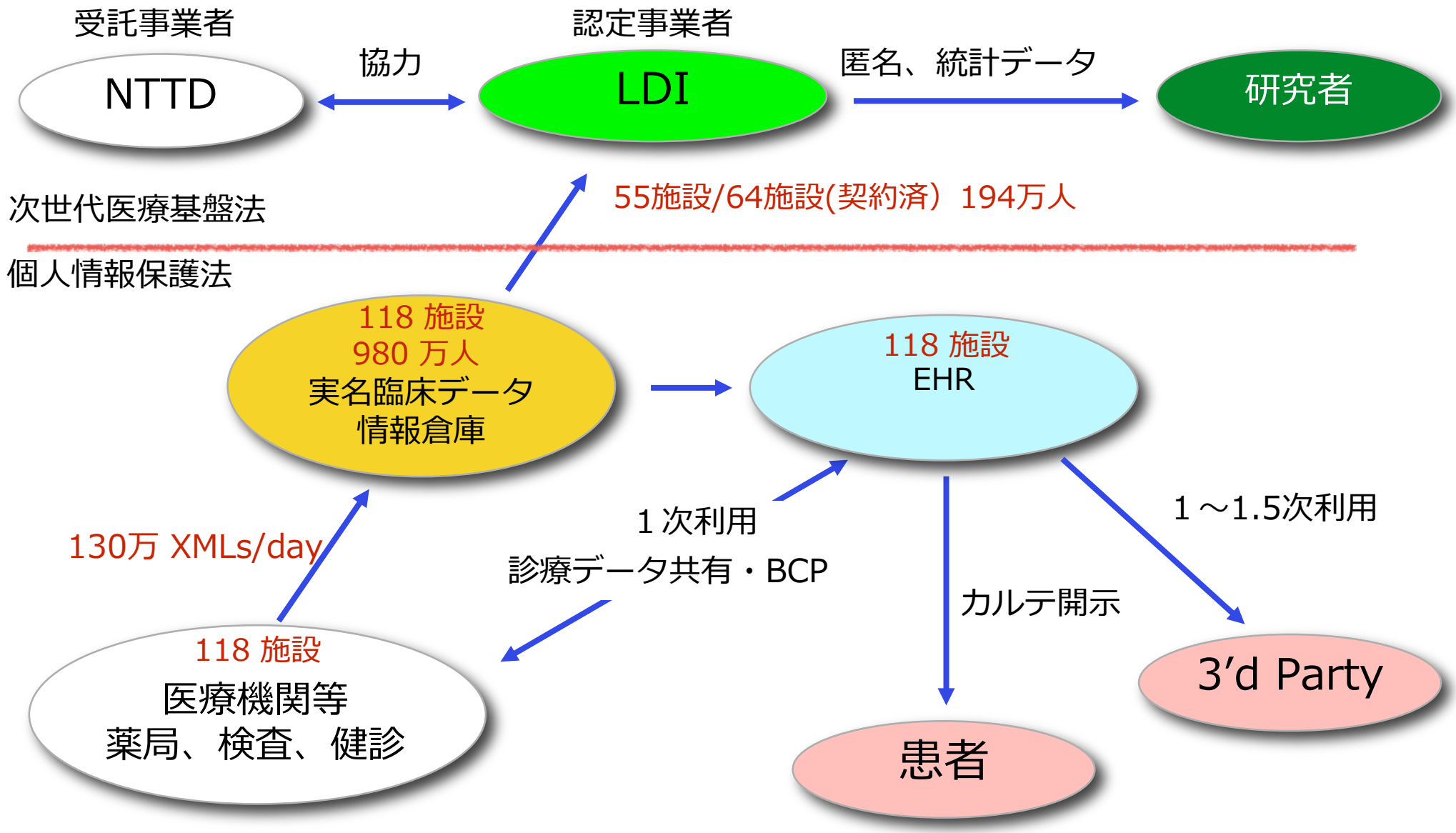


データ：冗長
モデル：分散
通信内容：活性
通信形式：AllReduce
通信頻度：層毎
長所：省メモリ
短所：通信オーバーヘッド
　　オーバーラップ不可
　　実装が複雑

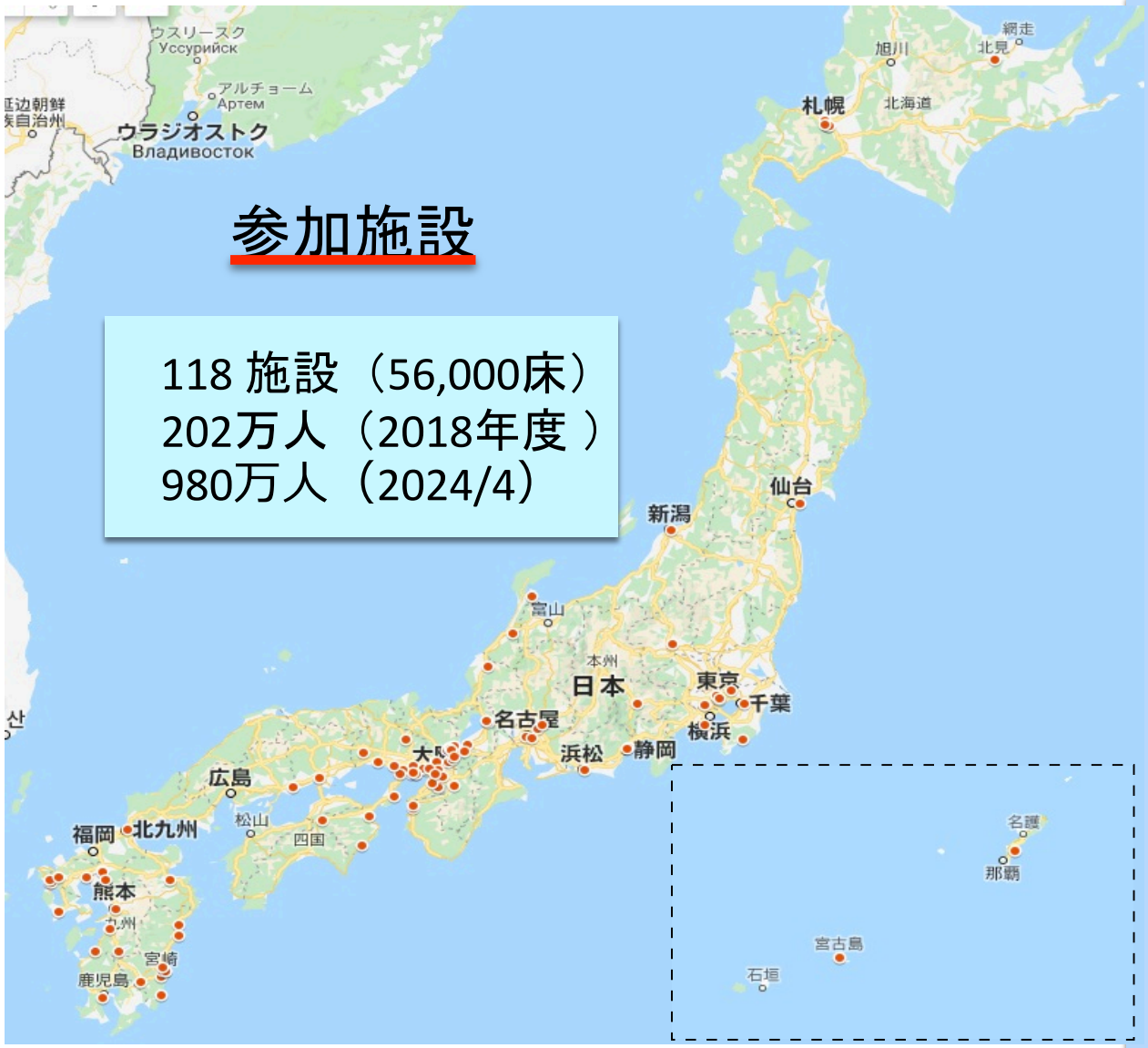
大規模電子カルテ学習データ：千年カルテ2024

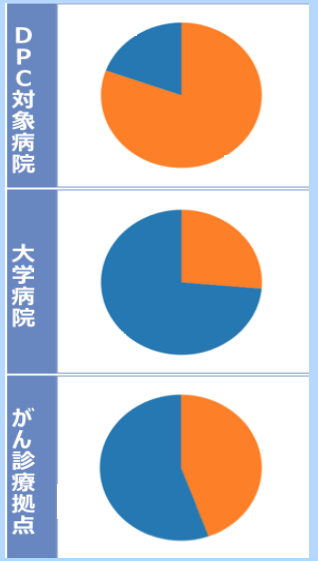
PRiME-R

2024.04 時点



千年カルテ：参加施設数

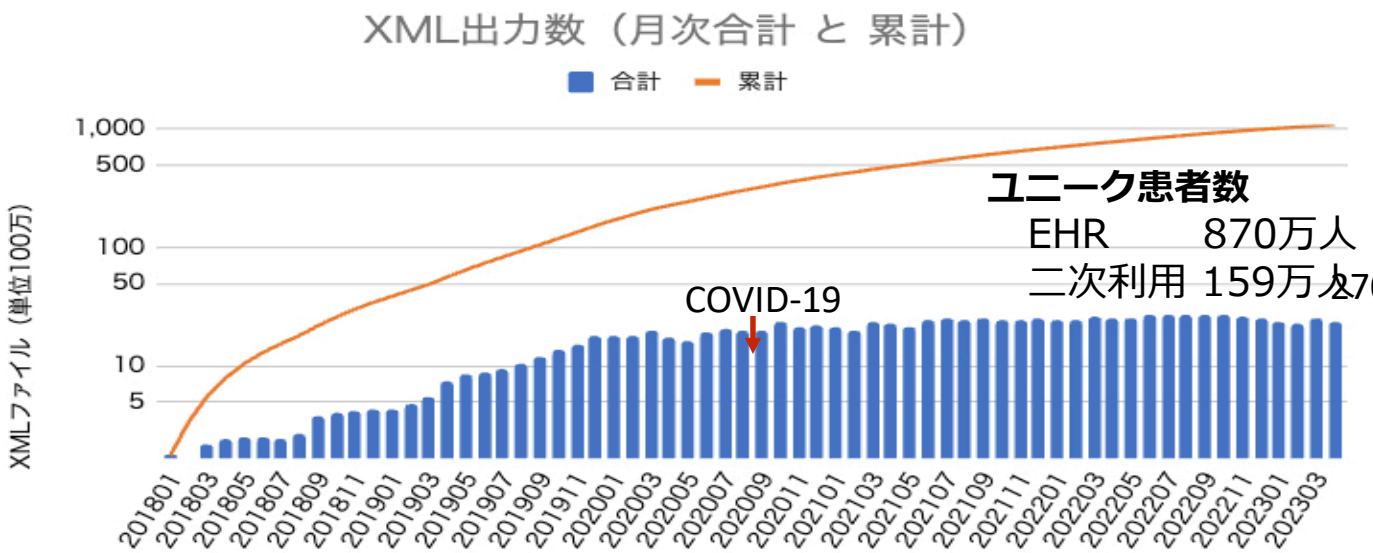

参加施設
118 施設（56,000床）
202万人（2018年度 ）
980万人（2024/4）
札幌
北海道
仙台
新潟
東京
千葉
横浜
名古屋
浜松
静岡
広島
福岡
北九州
熊本
鹿児島
宮崎
松山
四国
ウラジオストク
Владивосток
Уссурийск
Артем
旭川
北見
網走
富山
本州
日本
名護
那覇
宮古島
石垣


DPC対象病院
大学病院
がん診療拠点

千年カルテへのXMLデータ出力状況

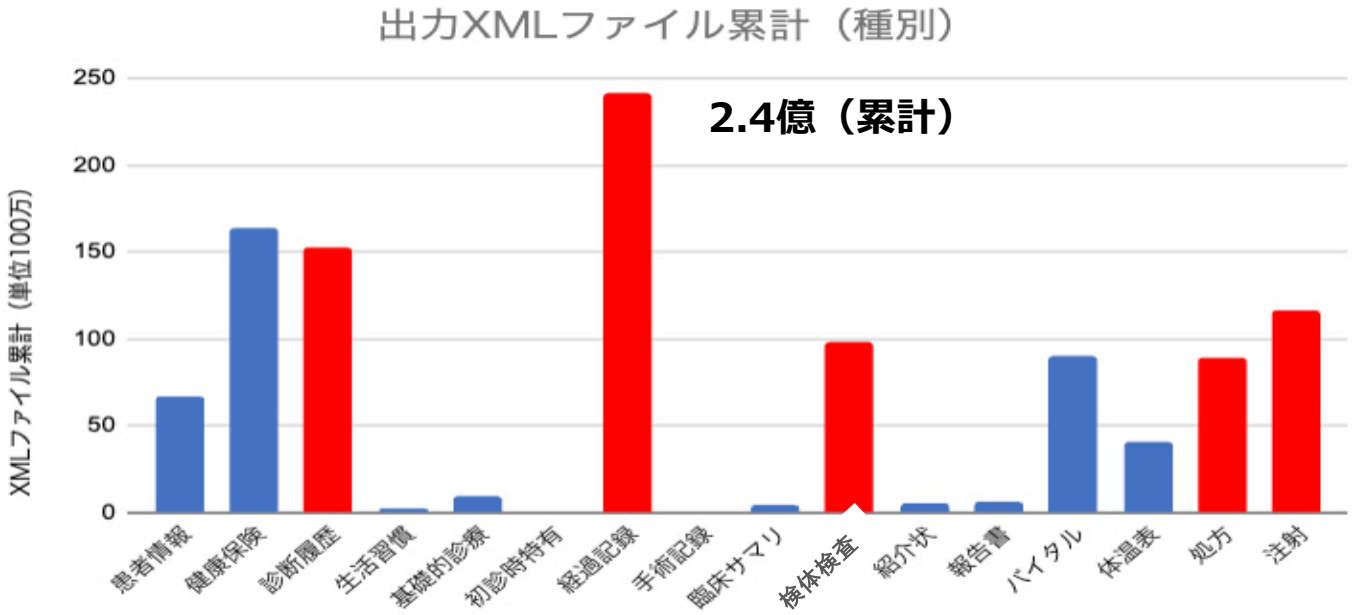
PRiME-R

千年カルテへのXMLデータ出力状況
2018.01～2023.04



11億（累計）

ユニーク患者数
EHR　870万人
二次利用 159万人 ※700万/月



2.4億（累計）