

令和5年度厚生労働省科学研究費補助金
循環器疾患等生活習慣病対策政策研究事業

「循環器疾患及び糖尿病、COPD等の生活習慣病の個人リスク及び集団リスクの評価ツール
の開発と応用のための研究（23FA1006）」
2023年度分担研究報告書

国際的なデータ標準に基づくコホートデータ統合の取り組み

研究協力者 筒井杏奈 東邦大学医学部社会医学講座医療統計学分野・助教
研究代表者 村上義孝 東邦大学医学部社会医学講座医療統計学分野・教授

研究要旨

本年度から開始したコホートデータの経時データの統合プロジェクトにおいて、臨床研究におけるデータベース構築の標準であるCDISC標準の取り決めに可能な限り活用しながら、質の高い統合データベースを効率的に作成した。実際に5コホートのデータを統合することで、CDISC標準を用いたデータ統合のポイントと問題点についてまとめた。その結果、CDISC標準群のうち、Analysis Data Model (ADaM) やCDISC Controlled Terminology等の一部が部分的に活用可能であることが明らかとなった。作成したデータベースを用い、予備的解析を行い、データ解析の実行可能性を示した。

A. 研究目的

EPOCH-JAPAN (Evidence for Cardiovascular Prevention from Observational Cohorts in Japan) は主に循環器疫学のコホート研究の個人単位のデータを収集・統合し解析する共同研究プロジェクトである。EPOCH-JAPANの参加コホートは17であり、高精度の曝露測定、対象者の高追跡率を誇る質の高いコホート研究データからなる IPD (Individual Participants data) Meta-analysis である。これまで EPOCH-JAPAN から公衆衛生政策と臨床診療ガイドラインの双方に資する科学的なエビデンスを提供してきた。本年度から新たな研究テーマである、経時データの統合プロジェクトが開始された。このプロジェクトは

EPOCH-JAPAN の構成メンバーの中のコホート研究で曝露の複数回測定のデータを有するコホートで声掛けをし、それらコホートの経時データおよびエンドポイント情報を収集・統合することで、健康・発症・死亡の全プロセスが解析できる新しい統合データベースを作成するものである。

統合データの作成に際してデータモデルの参照は重要である。データモデルとは、現実世界の対象をデータ集合として表現するための取り決めであり、近年、国際的に標準化が進んでいる。医学研究にて最も普及が進んでいるデータモデル標準の一つは、1997年設立の国際的で学際的な非営利団体「Clinical Data Interchange Standards Consortium

(CDISC)」が開発してきた標準群であり、CDISC 標準 (CDISC standards) と呼ばれる。これら標準は計画から収集、整理 (tabulation) および解析までの一連のプロセスを対象とし (図 1)、特に企業治験ではその実装が一般的になっている。その理由として主要な国々の薬事規制の下で新薬の承認申請時に、各 CDISC 標準に準拠した臨床試験の症例レベルのデータの提出が求められていることがあげられる。代表的な CDISC 標準として、データの整理あるいは解析に用いる Study Data Tabulation Model (SDTM) と Analysis Data Model (ADaM) が挙げられる。また Controlled Terminology (統制用語; CDISC CT) が管理されていることも特徴の一つである。これは米国国立がん研究所 (National Cancer Institute) の Enterprise Vocabulary Services (EVS) と協力して管理されているもので、SDTM および ADaM のデータの作成時に参照すべきルールとなっている。

本報告では共同コホートの経時的データを統合するため、CDISC 標準を可能な限り活用しながら、質の高い統合データベースを効率的に作成したので報告する。

B. 研究方法

(1) CDISC 標準の専門家に対するヒアリング

CDISC 標準を採用することの妥当性を検討する目的で、複数の CDISC 標準の専門家に、コホート研究の統合データベースに活用できるデータ標準についてヒア

リングを行った。

(2) データ統合プロセスの検討

CDISC 標準を採用する場合、一般に複数のデータの統合プロセスが必要となる。それらの統合プロセスを可視化した上で、本研究プロジェクトでも同様の手順を取るべきかについて、(1) のヒアリングを参考にして、実際に提供された 5 つのコホートデータの構造の検証を行った。

(3) データ変換と予備的解析

統合データベース作成の仕様書を作成したもとの、提供を受けた各コホートのデータを変換し、データを縦結合により統合した。その後、性別や年齢などについては予備的解析を行い、その実行可能性を確認した。データの変換と統合、および集計には SAS 9.4 を用いた。

(倫理面への配慮)

本研究計画は、東邦大学医学部倫理委員会から 2023 年 7 月 31 日に倫理委員会承認 (A23053) および 2023 年 11 月 9 日に研究機関追加の計画変更に対する倫理委員会承認 (A23075_A23053) を得た。

C. 研究結果

(1) データ標準の調査

CDISC 標準に関し、観察研究に特化したドラフト文書「Considerations and Examples of Using SDTM for Observational Studies - sections ~ Considerations and Examples for using SDTM for Observational Studies」を CDISC 標準の専門家より入手した。これは観察研究やリアルワールドデータに SDTM 標準を適用する際によ

く生じるイシュー (issue) を解決するための文書であり、応用例 (ユースケース) として症例対照研究やコホート研究が挙げられていた。ただこの文書は本研究プロジェクトで対象とする、健診などにより定期的に測定されるコホート研究に特化したガイダンスではなかった。

(2) データ統合プロセスの検討

本研究のデータ統合プロセス相談を受けた全て CDISC 標準の専門家は全員、図 2 や図 3 のようにまず各コホートデータを SDTM データに変換することを推奨した。その主な理由として、各コホートデータで異なるデータ構造や異なる統制用語 (例えば、男性の表現法も施設により「1」や「0」と異なる) を持つため、一度、統制用語も考慮しながら SDTM 標準に従って統一した方が、全体的なプロセスは効率化されて質が向上するということであった。実際、提供を受けた 5 つのコホートデータを確認すると、3 パターン (1 個のデータに全ての情報を格納したもの、ベースラインと経時データを分けて格納したもの、各年度の健診データと、追跡イベントを格納したデータ) に分けられた。1 つのコホートがもつデータセット数は最大で 6 と、一般の臨床試験データと比較すると少なかった。データ構造や統制用語が多種多様であった反面、被験者保護の観点から、どのコホートデータも研究に必要最小限の情報のみが提供されたために項目数は少なく、比較的シンプルなデータ構造が多かった。

以上のコホートデータの構造分析により、各コホートデータを SDTM 標準に従ってテーマ (トピック) ごとに多数のデータ

セットに分割することや、ADaM 標準に従ってデータを再統合することは労力が大きく、エラーが発生しやすいと予想された。したがって SDTM 標準への変換プロセスは省略することとした。

また ADaM 標準への適用可能性を考えると、まず一症例一レコードのベースラインデータは概ね Subject Level Analysis Dataset (ADSL) が適用可能と考えられた。他方、一症例複数レコードの経時データについては、今後、様々な検査値やアウトカム等の情報を用いて解析することが予想されていたため、垂直型の Basic Data Structure (BDS ; 表 1) を適用すると将来的にデータ修正が続発すると考えられた。それよりも、一症例一時点の構造で、変数の命名ルール (例: ベースラインの解析フラグは ABLFL と命名) ならびに統制用語 (例: フラグは「Y」(Yes の意) 等を格納) などのルールをできる限り採用して、BDS に似たデータを作成することとした。最終的に図 4 のように、各コホートデータは ADSL と ADREP データの 2 つの ADaM に類似したデータが作成された後、単純な縦結合をすることで統合する手順とした。

(3) データ変換と予備的解析

表 2 は作成された仕様書の一部を示す。統合データの構造を定義した上で、各コホートデータの変換方法を定義するマッピング仕様書を作成した。

その後、予備的な基本集計を行い、解析可能であることを確認した。この時、既存の臨床試験データで ADaM データを解析するために用いられた解析プログラムを流用することにより、少ないコードの修正で解析結果を作成することができた。

D. 考察

本研究では、国際的に普及している CDISC 標準の取り決めに可能な限り活用しながら、質の高い統合データベースの効率的な作成に取り組んだ。結果として、各コホートデータを、解析データモデルである ADaM 標準に似たデータに変換し、統合することで、国際的データモデルに可能な限り準拠した統合データベースを作成し、基本的な集計も可能であることが確認された。

標準化されたデータモデルを使う大きなメリットは、研究者あるいは解析担当者側からみると、データの調査にかかる時間を省略できることである。つまりデータの構造定義書等を見ずとも、一症例一レコードのデータは ADSL データに入っていることが予想できる。また症例の性別は変数 SEX に格納されて「M」の値は男性を指すことが予想できる。これにより、データの受領後はすぐ解析に取り組める。もう一つのメリットは、データ標準化により、既存の解析プログラムを必要最小限の修正で別の研究にも活用可能なことである。実際、本プロジェクトでも予備的な基礎集計にはそれほど多くの時間を要さなかった。またこの性質により、既存の解析プログラムの質が高ければ、本プロジェクトでの解析の質も高いことが期待される。

本研究の大きな限界点は、CDISC 標準は臨床試験を主たる対象としており、研究実施時点で観察研究に関するガイダンスがドラフト版で未公開だったように、観察研究はまだ開発・議論の余地がある点である。結果、本研究でも SDTM 標準を省

くなどの独自の形式を取らざるを得なかった。CDISC 標準以外の導入も検討すべきと考える。たとえば CDISC 標準以外では、Common data model (CDM) と呼ばれる、異なるアプリケーションやデータソース間でデータや情報を交換するための標準化されたデータモデルが存在していた。特に医学研究では、リアルワールドデータ (RWD) で保険請求データや電子カルテデータ等に基づき薬剤疫学的分析で用いられる Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) の普及が進んでいるようだった。特筆すべきは標準化ヴォキャブラリ (Standardized Vocabulary) が 800 万を超えていることで、CDISC の SDTM に対する統制用語が 4 万程度であることを考えると導入メリットがある。データ標準に関して今後も情報収集を進めたい。

E. 結論

本年度から開始したコホートデータの経時データの統合プロジェクトにおいて、CDISC 標準の取り決めに可能な限り活用しながら、質の高い統合データベースを効率的に作成した。実際に 5 コホートのデータを統合することで、CDISC 標準を用いたデータ統合のポイントと問題点についてまとめるとともに、予備的解析によりデータ解析の実行可能性を示した。

F. 健康危機情報

なし

G. 研究発表

1. 論文発表

なし

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

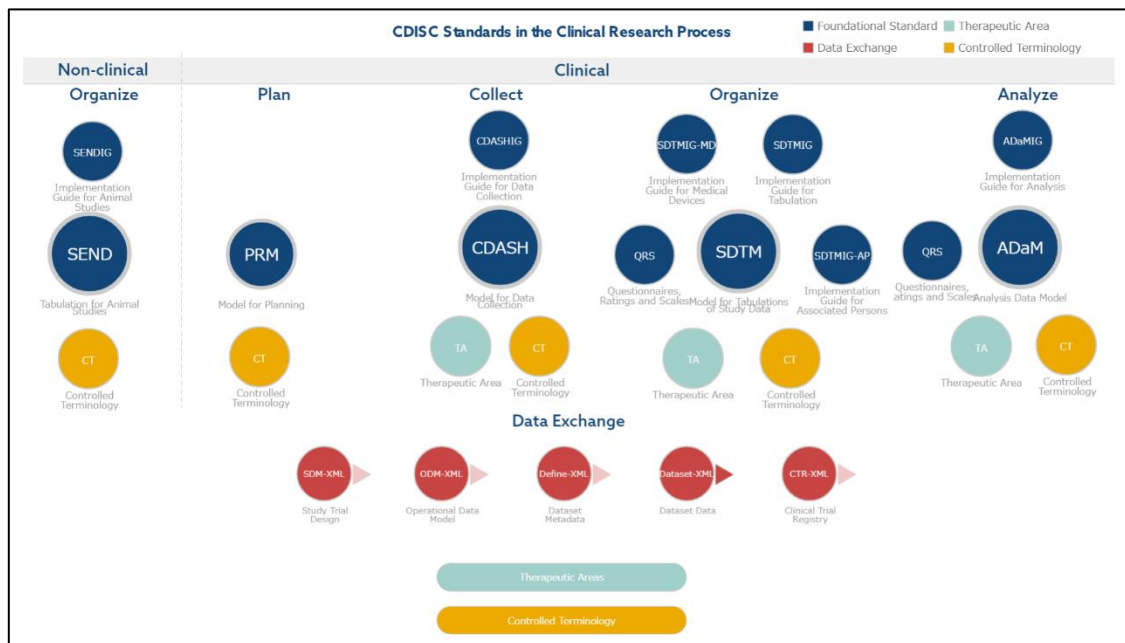


図1 臨床研究のプロセスにおける CDISC 標準

(CDISC 「Standards」 (<https://www.cdisc.org/ja/node/7026>) より一部改変)

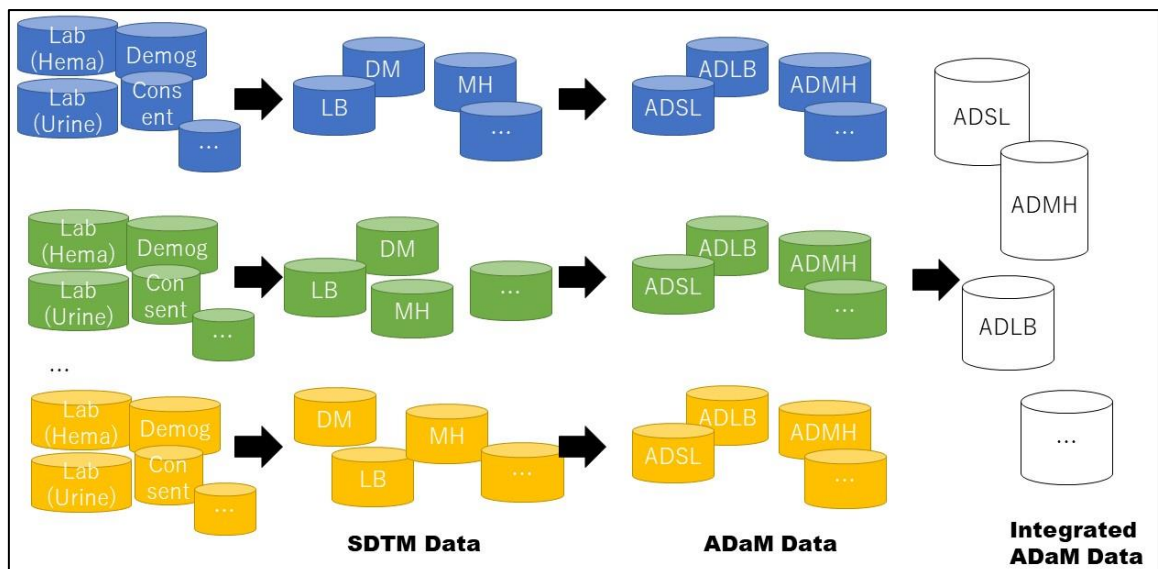


図2 臨床研究における一般的な統合データの作成プロセス①

AD, Analysis Data; ADSL, Subject Level Analysis Dataset; DM, Demographics; LB, Laboratory Test Results; MH, Medical history

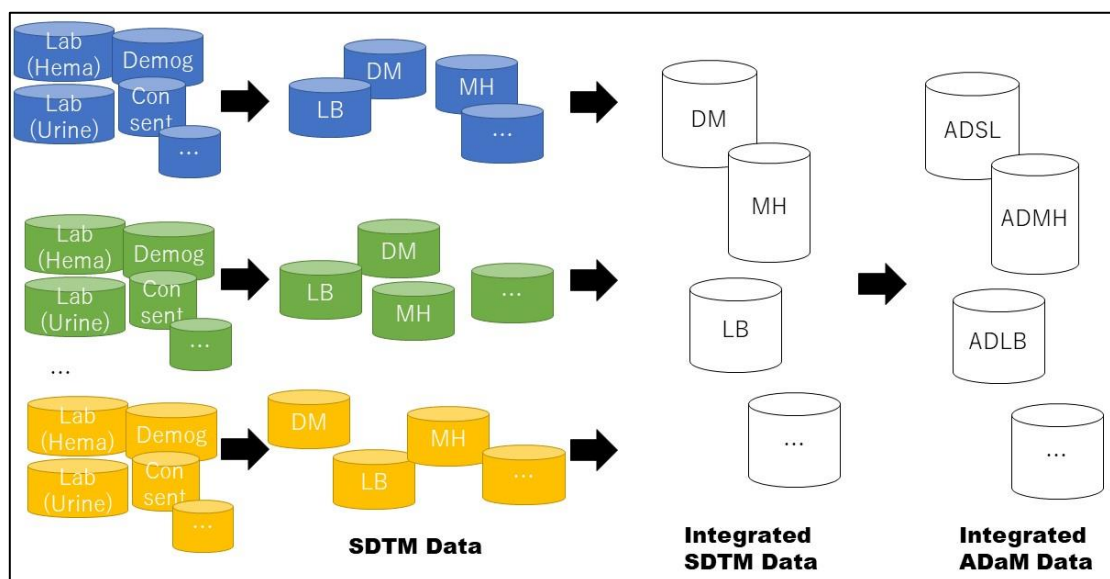


図3 臨床研究における一般的な統合データの作成プロセス②

AD, Analysis Data; ADSL, Subject Level Analysis Dataset; DM, Demographics; LB, Laboratory Test Results; MH, Medical history

Table 4.6.1. ADaM Dataset with Subject-Level and Record-Level Indicator Variables

Row	USUBJID	ITTFL	PPROTFL	VISIT	AVISIT	PARAMCD	AVAL	ANL01FL	PPROTFL
1	1001	Y	Y	Week 0	Week 0	TEST1	500	Y	Y
2	1001	Y	Y	Week 1	Week 1	TEST1	400	Y	Y
3	1001	Y	Y	Week 2	Week 2	TEST1	600	Y	Y
4	1002	Y	N	Week 0	Week 0	TEST1	500	Y	
5	1002	Y	N	Week 2	Week 1	TEST1	48	Y	
6	1002	Y	N	Week 2	Week 2	TEST1	46	Y	
7	1003	Y	Y	Week 0	Week 0	TEST1	999	Y	Y
8	1003	Y	Y	Week 1	Week 1	TEST1	999		Y
9	1003	Y	Y	Retest	Week 1	TEST1	49	Y	

表1 CDISC ADaM 標準の BDS データの一例

(CDISC Analysis Data Model Team 「Analysis Data Model Implementation Guide Version 1.3 (Final)」 の Table 4.6.1 より抜粋)

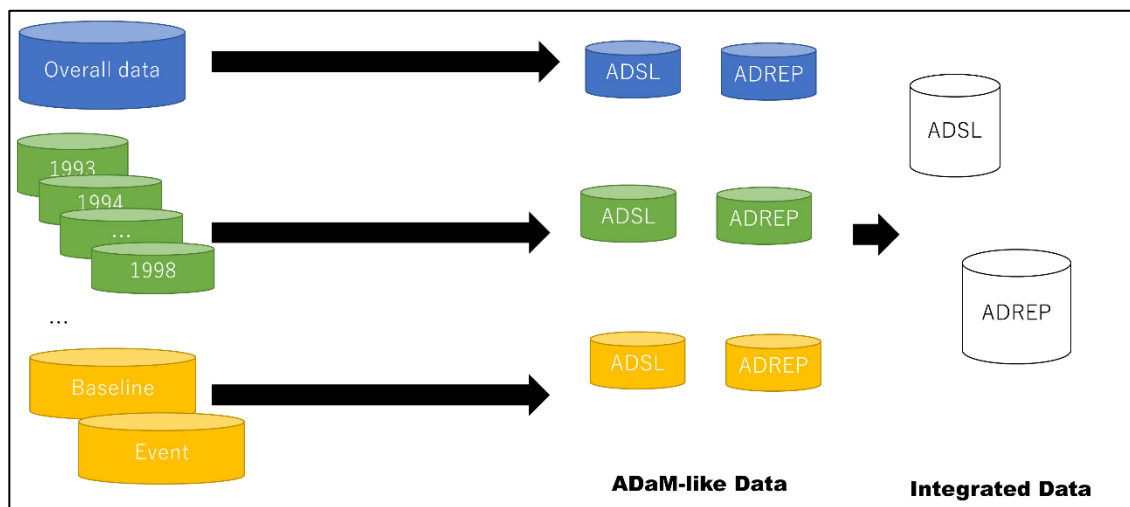


図4 本研究で活用したプロセス

AD, Analysis Data; ADSL, Subject Level Analysis Dataset; DM, Demographics; LB, Laboratory Test Results; MH, Medical history

A	B	C	G	H	S
大項目	小項目	ラベル名 (BL*には"Baseline"を付与)	変数名 (BL)	変数名 (経時)	共通仕様または説明
1					
2					
3	IDなど				
4		Study Identifier	STUDYID	STUDYID	管理番号+コホート名
5		Unique Subject Identifier	USUBJID	USUBJID	=catx("/", STUDYID, SUBJID) 例: 12RERF/00001
6	ID, 健康診断日など ソートに用いる変数	ID	SUBJID	SUBJID	各コホート内のID
10	(健診日)	Study Date of Examination		EXDTC	
11		Study Day of Examination		EXDY	EDAT - STARTDT, If EDAT >= STARTDT then +1。ペ
14	データ提供依頼に含まれ				
15		性別	SEX	SEX	M / F
16		Sex (N)	SEXN	SEXN	1 / 2. IFN(SEX="M", 1, IFN(SEX="F", 2, .)) ;
17	年齢	Age (Years)	BLAGE	AGE	
18	身長	Height (cm)	BLHEIGHT	HEIGHT	
19	体重	Weight (kg)	BLWEIGHT	WEIGHT	
20	BMI	BMI (kg/m2)	BLBMI	BMI	
21	喫煙 (非喫煙、禁煙、現在喫煙)	Smoking	SMOKE		NEVER / FORMER / CURRENT
22		Smoking (N)	SMOKEN		1 / 2 / 3. IFN(SMOKE="NEVER", 1, IFN(SMOKE="FORMER", 2, 1
23	飲酒 (非飲酒、禁酒、現在飲酒)	Alcohol	ALCOHOL		NEVER / FORMER / CURRENT
24		Alcohol (N)	ALCOHOLN		1 / 2 / 3. SMOKENと同じ導出方法
30	収縮期血圧	Systolic Blood Pressure (mmHg)	BLSYBP	SYBP	
31	拡張期血圧	Diastolic Blood Pressure (mmHg)	BLDIABP	DIABP	
32	総コレステロール (注: 測定間隔が2年に 一度の場合は、3回分(6 年間)とする)	Total Cholesterol (mg/dL)	BLCHOL	CHOL	
33	HDL-C	HDL Cholesterol (mg/dL)	BLHDL	HDL	提供データ依頼より追加
34	Non-HDL	Non-HDL Cholesterol (mg/dL)	BLNONHDL	NONHDL	
35	ヘモグロビンA1c	Hemoglobin A1C (%)	BLHBA1C	HBA1C	JDS値
38	服薬状況-糖尿病	Concomitant Medications - Diabetes	BLCMDM	CMDM	Y / N

表2 経時コホートデータベースにおける仕様書 (一部)

付録 図2, 図3に基づくデータベース構築に関する説明

臨床研究で CDISC 標準を採用する場合、収集データの主な統合プロセスは図2または図3であらわされる。両プロセスとも、最初にデータが SDTM 標準に従って症例一覧データ (SDTM データ) に変換される。この時、被験者背景データ (DM, Demographics) や臨床検査値データ (LB, Laboratory Test Results) 等、データのテーマ (トピック) ごとにデータは分割される。その後のプロセスは図2と図3で異なる。

図2であれば ADaM 標準に従って解析データ (ADaM データ) に変換される。ADaM 標準では、まず Subject Level Analysis Dataset (ADSL) と呼ばれる、一症例につき一レコード (one record per subject) のデータが作成される。次に、Basic Data Structure (BDS) と呼ばれる、一症例一解析パラメーター一解析時点につきまたは複数レコード (1 or more records per subject, per analysis parameter, per analysis timepoint) のデータが作成されることが多い (図4)。データ構造は、解析に用いる値を変数名 AVALC 等の決められた変数に格納する垂直型となっており、どの BDS データの解析でも同じ解析であればほぼ同じ解析プログラムが使えるようになっている。また、原則として「そのまま解析が可能 (Analysis-Ready)」がある。例えば解析で収集項目にない推算糸球体濾過量 (eGFR) の集計が必要であれば、SDTM データ作成に当たり分割された情報、すなわち DM データに含まれる年齢及び性別ならびに LB データに含まれる血清クレアチニン値を改めて結合した上で eGFR 値を導出し、ADLB データに格納させることが必要となる。多数の処理を経て ADaM 標準に従って全ての ADaM データの作成された後は、各データの統合プロセスに入る (図2の Integrated ADaM)。これまでのプロセスでデータが標準化されているため、同じ名称の各データの単純な縦結合で統合が完了する。

図3のプロセスであれば、各臨床試験データの SDTM データが作成されたのち、それらを縦結合により統合される (Integrated SDTM データ)。また必要に応じて、さらに解析用の統合データ (Integrated ADaM データ) が作成される。