第27回日本医療情報学会春季学術大会
チュートリアル6

二次利用ユーザーのためのデータ抽出、変換、登録（ETL）入門

－ETLの理解がデータ活用の近道－

一般社団法人SDMコンソーシアム

アジェンダ

資料
https://sdm-c.org/report.html

はじめに
千葉大学医学部附属病院　本多　正幸（ほんだ　まさゆき）

ETLプロセスの概要
SDMコンソーシアム　鈴木　英夫（すずき　ひでお）

ETL開発の実際
熊本大学病院　山ノ内　祥訓（やまのうち　よしのり）

PHRデータ利活用事始め（Apple Healthcareからのデータ取得と医療情報との連携）
東京医科歯科大学　内村　祐之（うちむら　ゆうじ）

病院内のデータマネジメント・利活用を目的としたETLの実績と考察
蒲郡市民病院　飯田　征昌（いいだ　まさよし）

第27回日本医療情報学会春季学術大会
チュートリアル6

二次利用ユーザーのためのデータ抽出、変換、登録（ETL）入門
－ETLの理解がデータ活用の近道－

ETLプロセスの概要

一般社団法人SDMコンソーシアム
鈴木英夫

第27回日本医療情報学会春季学術大会
ＣＯＩ開示

演題名：　ETLプロセスの概要

筆頭演者名：　鈴木　英夫

私が発表する今回の演題について開示すべきＣＯＩはありません。

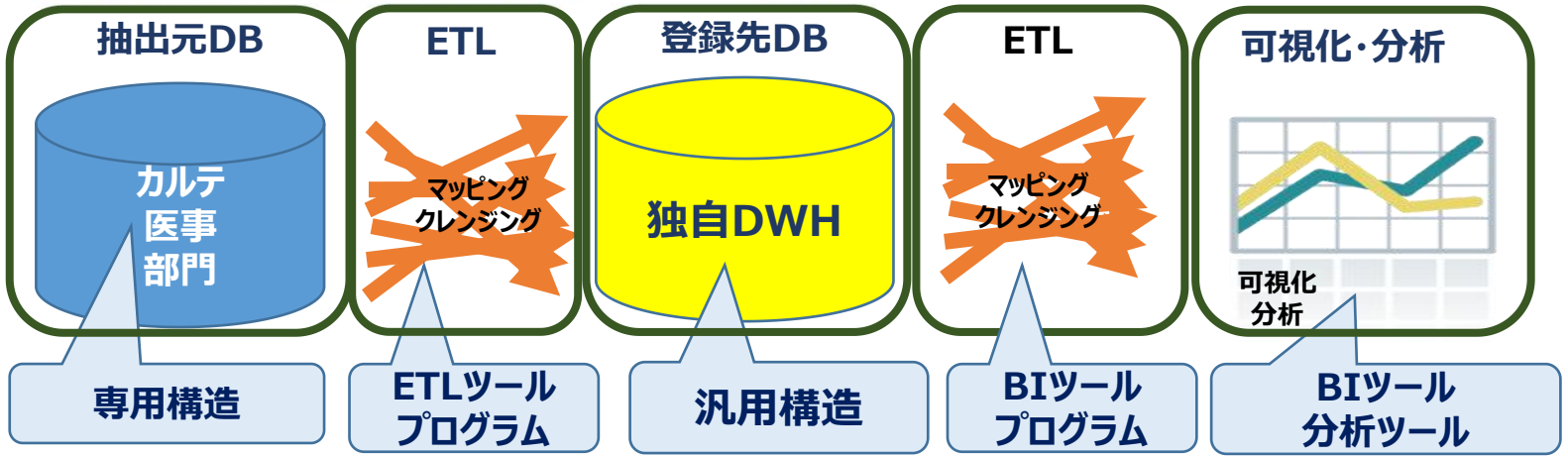
ETLプロセスの概要

- 二次利用におけるETL
  - ETLの位置づけ
  - ETL開発の事前チェック項目
- ETL開発プロセス
  - 設計
    - テーブルマッピング
    - 抽出条件
    - エンティティ・マッピング
  - 開発
    - オフィスツールの利用
    - 専用ツールの利用
    - プログラム開発

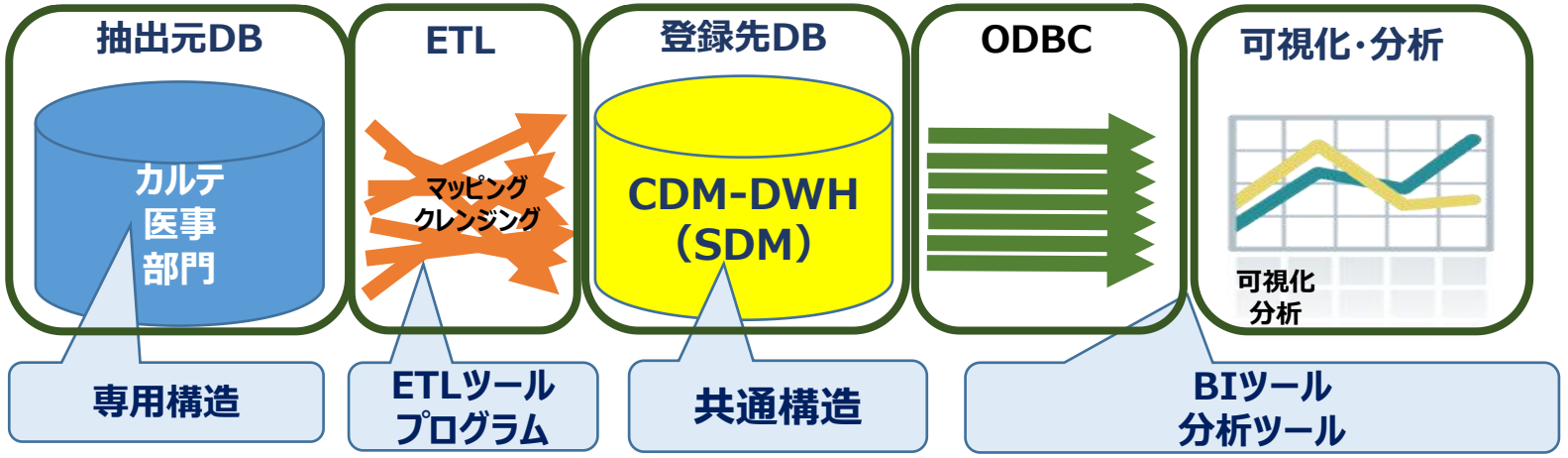
- ETL開発プロセス（続き）
  - テスト
    - テストデータによる単体テスト
    - タスク実行による統合テスト
    - 実環境による負荷テスト
  - 検証
    - 設計との比較
    - 抽出元との比較
  - データ移行
    - 移行方法
  - 運用・保守
    - 管理ソフト
    - 問題対応

ETL（Extraction Transformation Load)の位置づけ

医療情報の二次利用（独自DWHの場合）



医療情報の二次利用（共通モデルDWHを用いる場合）



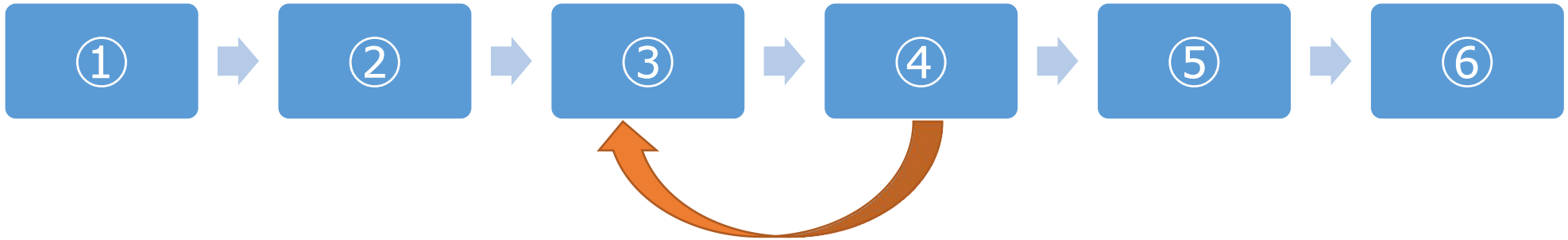
SDM Consortium
®登録商標第6025526号

ETL開発における事前チェック項目

抽出元からのデータ取得方法
- データベースとの接続　（RDB、DWH、VIEW）
- オフライン接続（CSV、XML、JSON）
- PUSH方式、PULL方式、非同期、手動

- データ定義書
  - テーブル定義、ER図
  - 守秘契約
  - 費用
- システム要件
  - インフラ　（仮想、物理、DBサーバ、アプリサーバ）
  - DB　　（商用DB、オープンソース）
  - ネットワーク　（共用、専用）
- 体制
  - 要求元担当（打ち合わせ、ベンダー間調整、院内調整、入退室管理）
  - 抽出元担当（開発、Q＆A）
  - ETL開発担当（課題管理、作業報告）

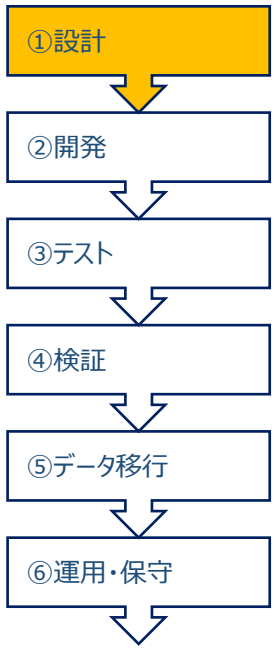
ETL開発プロセス



ETLの開発プロセス
①設計：（テーブルマッピング）（抽出条件）（エンティティマッピング）
②開発：（オフィスツールの利用）（専用ツールの利用）（プログラム開発）
③テスト：（テストデータによる単体テスト）（タスク実行による統合テスト）（実環境による負荷テスト）
④検証：（設計との比較）（抽出元との比較）（修正）
⑤データ移行：（移行方法）
⑥運用・保守：（管理ソフト）（問題対応）

ETL開発プロセス：設計

リレーショナルデータベース（RDB）は、最も普及しているDBで
市場占有率７０％を越えているので、RDB直接接続を前提とする



テーブルマッピング

テーブルマッピング
抽出元と登録先のテーブル間の対応付け
（例）　抽出元：処方オーダ　→　登録先：処方オーダ

抽出条件

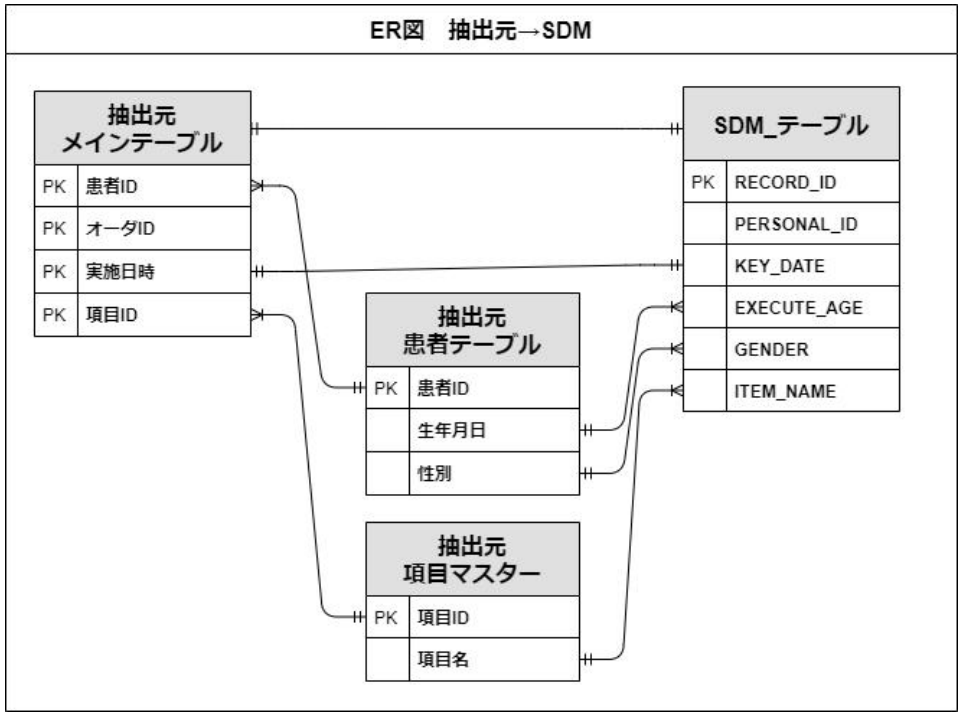
抽出条件
抽出元テーブルのレコード（行）単位と登録先のレコード（行）単位の関係
（例）　処方：同一オーダ、RP、薬品の用法単位　→　薬品単位

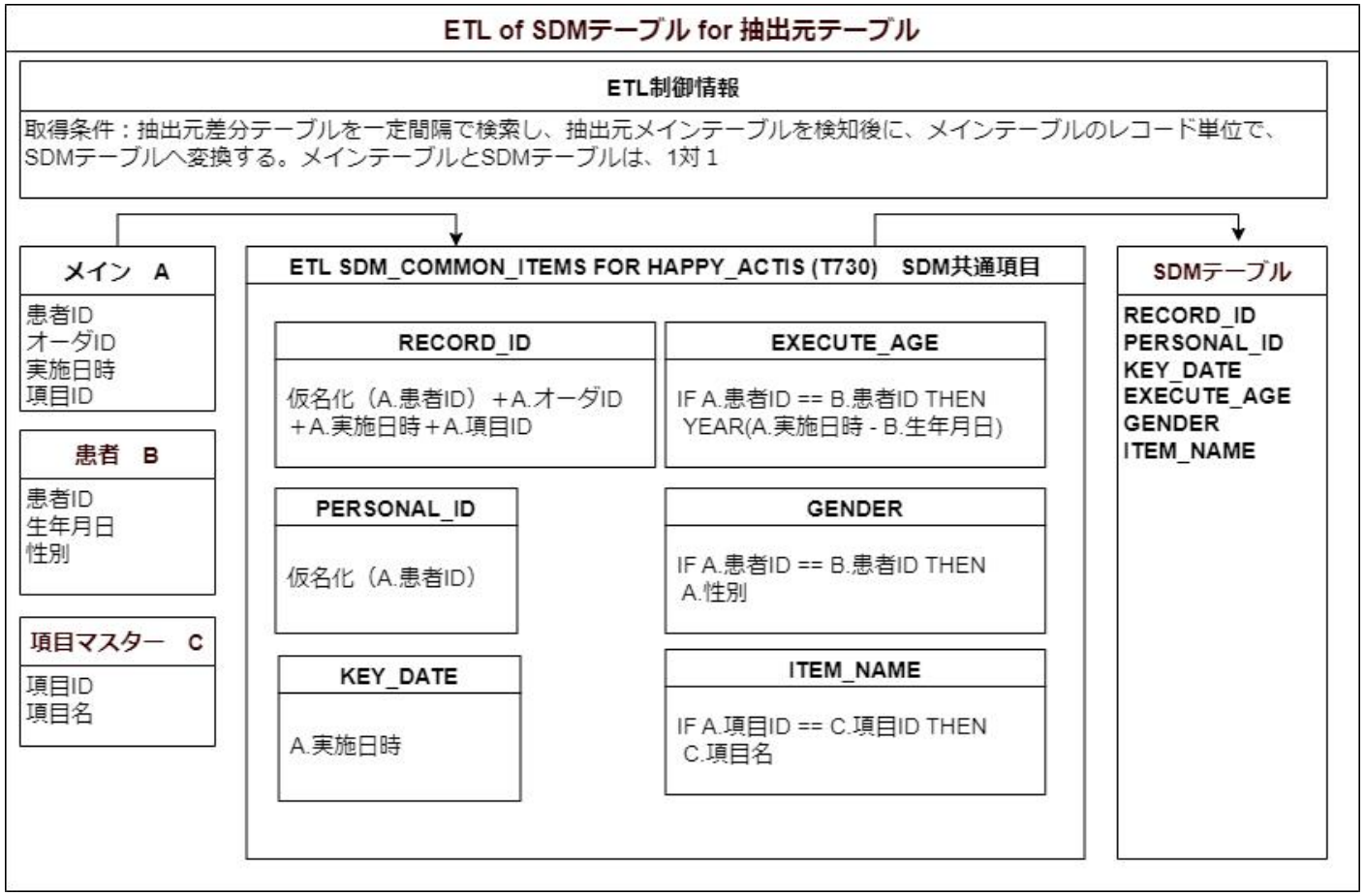
エンティティ・マッピング

エンティティ・マッピング
抽出元テーブルのエンティティと登録先テーブルのエンティティの対応付け
（例）　RP番号　→　RP_ID

RDBの用語
スキーマ、SQL、キー、インデックス

ETL開発プロセス：設計の例





ETL開発プロセス：開発

①設計
②開発
③テスト
④検証
⑤データ移行
⑥運用・保守

オフィス、BIツール

OR

専用ツール

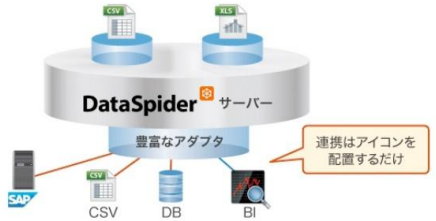
OR

プログラム開発

オフィス、BIツール
MS ACCESS、MS EXCELなどの組み合わせ
BIツールに組み込まれているETL機能

専用ツール（商用）
ETL専用ツール：DataSpiderなど
設計、実行、スケジューラなどオールインワン

さまざまなファイル形式、DB、BIツールのデータを
直感操作でひとまとめ

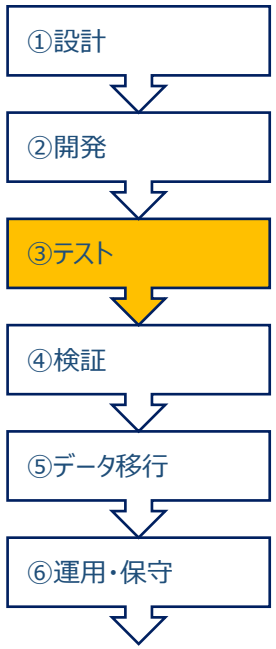


出典：大塚商会HP

プログラム開発
.NET (C#,VB,F#)、Java、Pythonなど

推奨構成（低コスト）
DBサーバー：Linux
アプリサーバー：Windows
DB：PostgreSQL
言語：Python

ETL開発プロセス：テスト



単体テスト

統合テスト

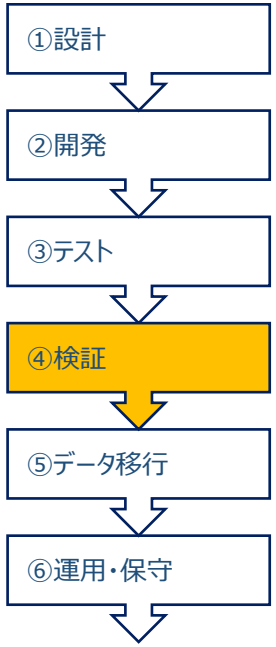
負荷テスト

単体テスト
登録先テーブル単位のETLでのテスト

統合テスト
登録先全テーブルETLタスクでのテスト

負荷テスト
タスクスケジューラによる日次テスト

ETL開発プロセス：検証



設計との比較

抽出元との比較

修正

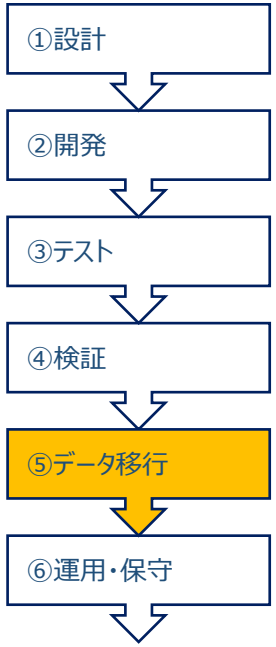
設計との比較
登録先テーブルのマッピングされている項目にデータが入っているか
抽出元と登録先の同一項目（マッピングされている項目）に同じ値が入っているか

抽出元との比較
一定期間（例：１年）の抽出元、登録先の重複なし患者数が合っているか
合っていない場合、その原因を調べる
（延べ患者数（レコード件数）は、ETL設計上合わないことがあるため）

修正
設計との不一致や、マッピングミスなどの修正を行い、単体テストを行う

検証は、SQLで行っても良いが、BIツールを用いる方が効率が良い

ETL開発プロセス：データ移行



カレンダー順

過去から現在（昇順）の移行
同一項目が上書きされる場合（患者プロファイルなど）降順で移行すると最新のデータを取得できない場合があるため
日時は、発生日でINDEXが付与されている項目を選択する

OR

レコード順

レコードの昇順での移行
DBが生成するレコードIDがある場合は、並べ替えが不要なため、高速に抽出できる

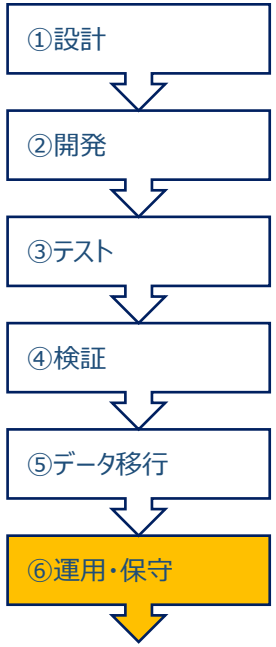
OR

ID順

ID順での移行
患者IDなど、必須項目かつプライマリキーとなっている場合、ID単位でかつ昇順に移行すると高速に移行可能となる

SQLのGROUP BYやORDER BYは、抽出の処理時間がかかるため可能な限り移行には用いない

ETL開発プロセス：運用・保守



管理ソフト

問題検知

問題対応

管理ソフト
　ETL処理のログを管理するソフトを作成する。通常のログおよびエラーログを分けることにより、ユーザーは正常かどうかの判断が可能となる

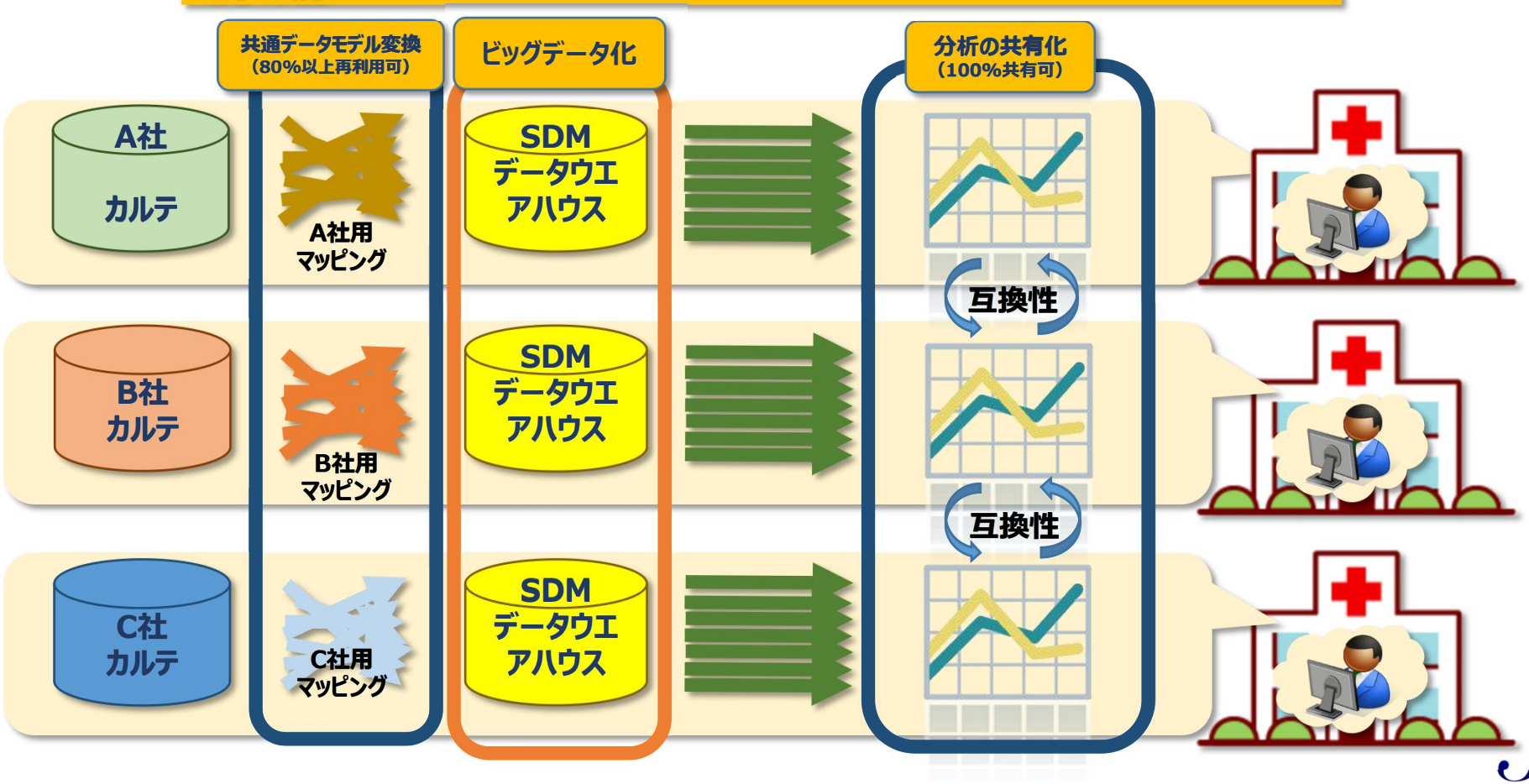
問題検知
　正常処理のログが記録されていない場合、およびエラーログが記録された場合に、開発元に連絡する

問題対応
　問題に対する対応が指示された場合、指示通りに実行し、問題発生日から、ETLの再実行を行う

指示に登録済みのデータの一部を消去する内容がある場合、実行中のETLを停止すべきかどうかも確認する

共通データモデルによりビッグデータ分析が可能

SDMは定義が統一されているため、分析手法を共有することが可能、また患者IDを匿名化しているため各所のデータを収集し容易にビッグデータ化することも可能



SDM Consortium

®登録商標第6025526号

SDMは、イベント記録ではなく、行為を記録するモデルです

SDMは共通項目により、いつからいつまで、どこで、誰が、何の役割で、どこの誰に対して、何の行為を行ったかを表現するようにしています。この共通項目は、検索や集計の軸として、また他テーブルとの結合に利用できます。

共通項目(英語)	共通項目(日本語)	項目の内容
RECORD_ID	レコードID	レコードを特定するキー、KEY_DATEを含むテーブル毎にSDM項目から設定
GROUP_ID	グループID	グループを特定するキー
TRANSACTION_ID	トランザクションID	トランザクションを特定するキー、TABLE_IDを含むテーブル毎にSDM項目から設定
RECORD_KEY	レコードキー	レコードを特定するキー項目名(SDM側)をカンマで連結
GROUP_KEY	グループキー	グループを特定するキー項目名(SDM側)をカンマで連結
TRANSACTION_KEY	トランザクションキー	トランザクションを特定するキー項目名(SDM側)をカンマで連
PERSONAL_ID	個人ID	個人を特定するIDの暗号化情報
INTEGRATE_ID	名寄せID	同一人PERSONAL_IDの最小のものに統一
ORGANIZATION_CODE	施設コード	施設コード
ORGANIZATION_NAME	施設名	施設名
KEY_DATE_TYPE	キー日時の種別	絞り込み対象日時の種別(テーブル毎に設定)
KEY_DATE	キー日時	タイムスタンプ型で絞り込み対象の日時を入れる
ACTION_TYPE	行為タイプ	入院、外来、健診、訪問、入居、治験、その他
DIVISION_CODE	発生場所コード	病棟、フロア、部屋などのコード
DIVISION	発生場所のアドレス	病棟、フロア、部屋など
SECTION_	部署コード	診療科、部署コード
SECTIO	部署	診療科、部署
REC	登録日時	初回記載日　タイムスタンプ型
UP	修正日時	変更日　タイムスタンプ型
STOP_REASON	停止理由	停止の理由
FIXED_DATE_TYPE	確定日時の種別	確定日の種別
FIXED_DATE	確定日時	確定日　タイムスタンプ型
AUTHOR_TYPE	著作者の種別	著作者の種別
AUTHOR_ID	著作者ID	著作者ID
AUTHOR	著作者	著作者
AUTHOR_OCCUPATION	著作者の職種	著作者の職種
RECORDER_ID	記録者ID	起票者のID
RECORDER_NAME	記録者	起票者の名前
RECORDER_OCCUPATION	記録者職種	起票者の職種
TEAM_TYPE	チームの種別	チームの種別
TEAM_ID	チームID	チームID
TEAM_NAME	チーム	チーム名
TEAM_MEMBER_ID	チームメンバーID	チームメンバーID(カンマで連結)
TEAM_MEMBER_NAMES	チームメンバー	所属:メンバー　(カンマで連結)
ORIGINAL_RECORD_KEY	オリジナルレコードキー	ANDで連結　項目名=データ
LOAD_TIMESTAMP	登録日時	SDMテーブルへの登録日時
EXPIRE_TIMESTAMP	有効日時	最新データの場合、9999/12/31 23:59:59.999を設定し、最新データが挿入された場合に、そちらのKEY_DATEに置換するKEY_DATEが小さいものが挿入された場合に、小さい方は自身のKEY_DATEを挿入し、当レコードは変更なし

患者IDではなく匿名化ID（他DWHは患者ID)

FOR　WHOM　誰に対して

WHEN　絞り込みに用いるキー

WHERE　物理的な場所

WHERE　論理的な組織

WHO　レコードの責任者

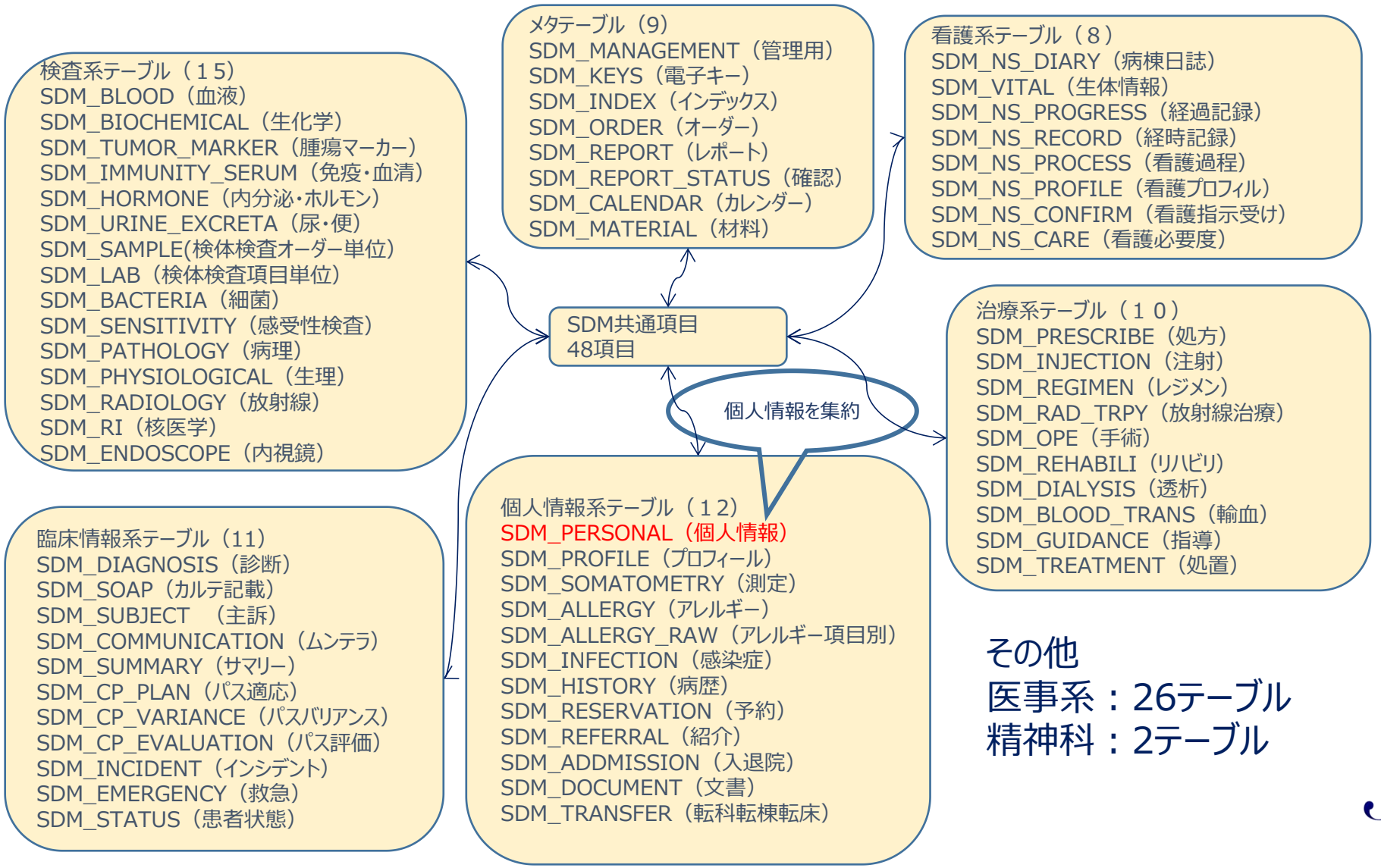
WHO　レコードの記録者

WHO　行為のチーム

最新レコードの特定情報

V1.12より追加した共通項目
承認者 Autholizer
開始時刻　BEGIN
終了時刻 END
所要時間 DURATION
検索日 QUERY_DATE

SDMテーブルV1.12（92テーブル、8170項目）

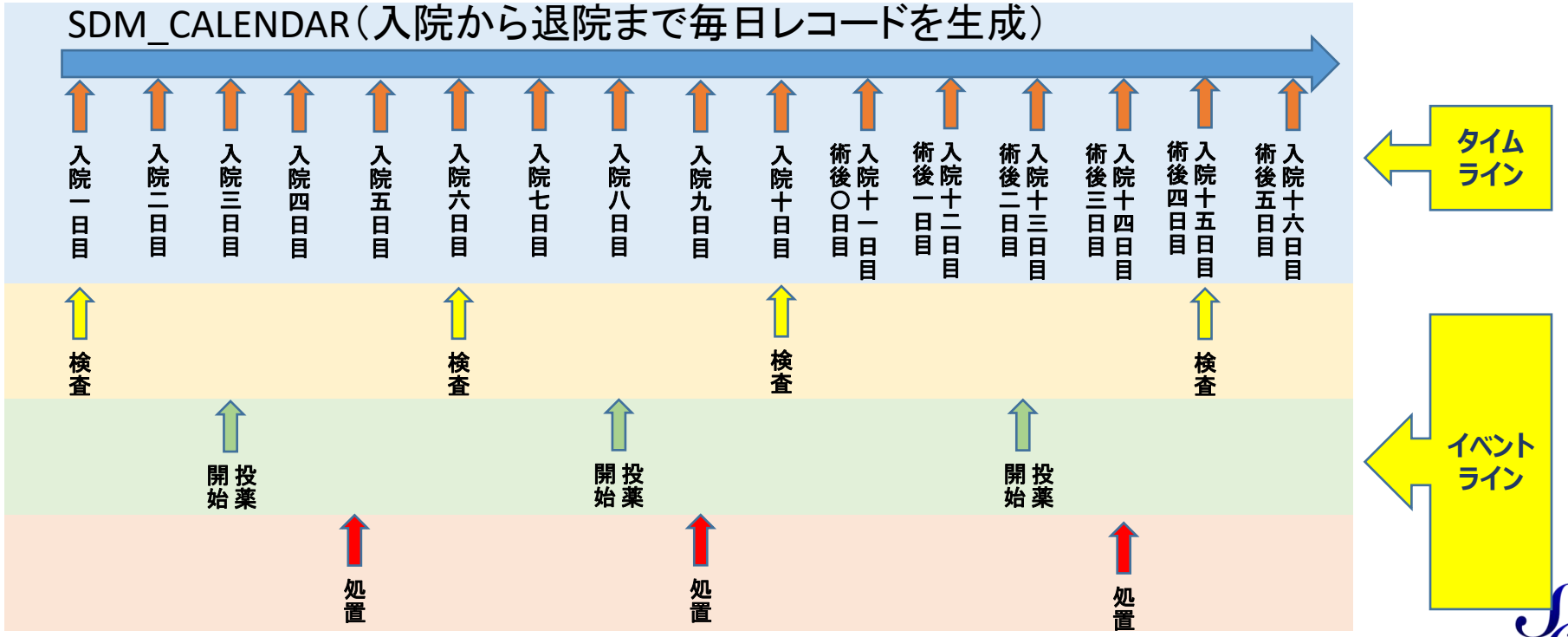


その他
医事系：26テーブル
精神科：2テーブル

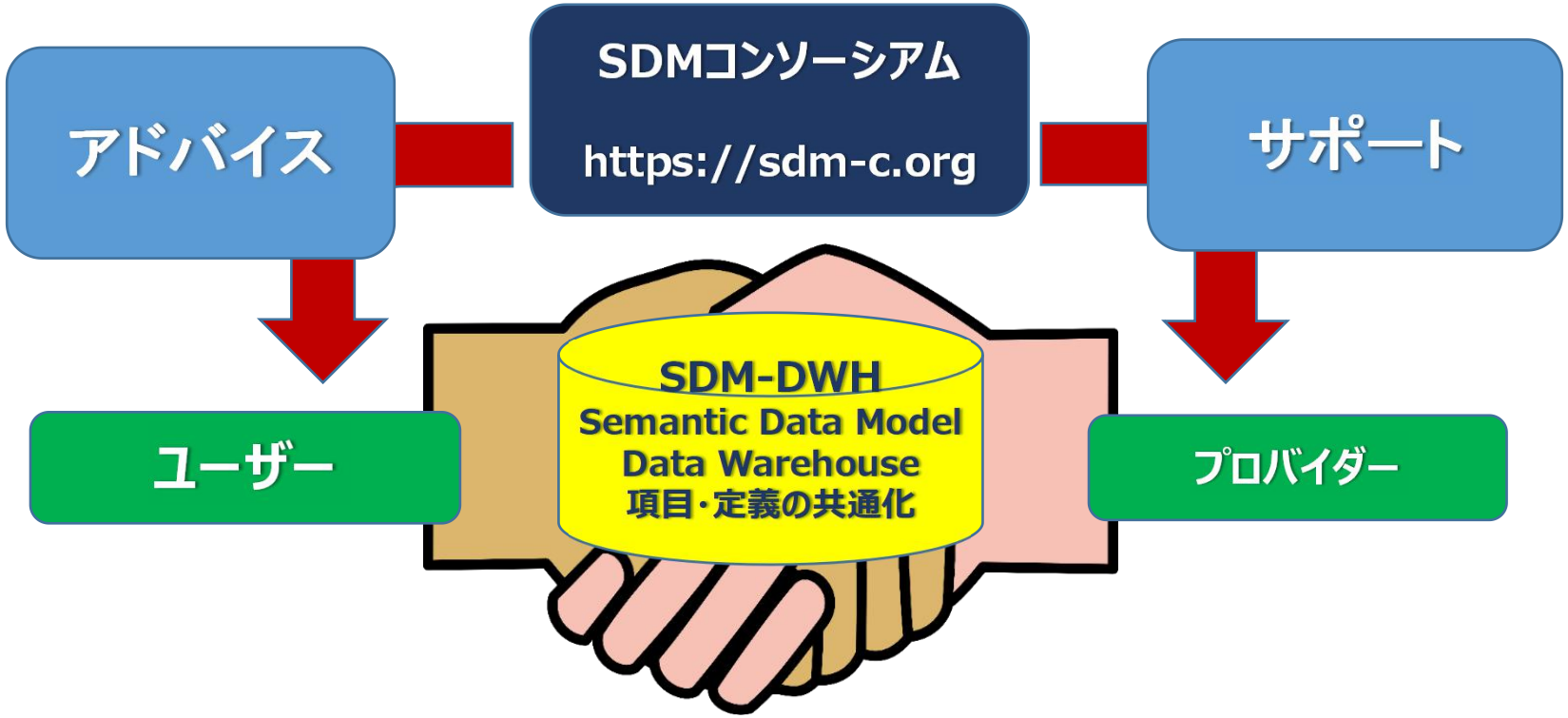
タイムライン対応テーブル

SDM_CALENDAR（入院カレンダー）は、時間軸として利用可能

- 電子カルテなどのほとんどはイベントベースのテーブルであり、複数のテーブルを結合する際、同一日が存在しないため結合できないことがあるため、投薬何日目から効果があったかを知ることが難しい
- SDM_CALENDARは、入院から退院までの毎日、必ず相対日が記録されるため、すべてのイベントが、入院何日目か、手術何日目かを知ることが出来る



SDMコンソーシアムの役割



コミュニケーション	渉外活動	広報活動	SDM規格管理
研究・教育・研修	イベント企画	スキル育成	運営・事務管理

SDMコンソーシアム・プロファイル
https://sdm-c.org/profile.html
第９期　役員（１０名）賛助会員：（１３社）
理事会、運営委員会、広報委員会、品質向上委員会

定義書

SDM Consortium