○ 返戻がかかっていることが分かるフラグが欲しい。データセットに入れるかどうかは研究者が決める。

○ 返戻は特定の疾病や医療機関に偏る可能性もあり
全ての返戻を除いた場合、どれだけ違うか検証してみては

特定健診の事例の有無

○ 特定健診の事例は？
（厚労省）模擬審査で1件あり。
申し出があっても採択されず

抽出条件からデータ量の予測がつかない

○ 抽出条件を指定して、まず概略の件数が分からないか？　データ量の大小で条件を変えて申請したい
（厚労省）プロセスとしてない
抽出条件次第では、研究に必要とされる以上に大きなデータになってしまっている可能性は？

抽出方法(言語:SQLの検討)

○ 研究者のニーズをSQLで受け取るのはどうか？
（富士電機）SQLだと半日で済む、オペレータの介在時間が減る
現状は、あいまいな仕様確認に時間がかかっている

○ 研究者にとって、今のレセプトデータ仕様ではSQL文を作るのは難しい。中間モデルを作るとしても、今のデータ格納方法が良いのか？

申出者のレセプト理解

○ 今までの申し出はレセプトを扱ったことがある人が多かった。
しかし、その中にはどうみても変な申し出もあり、この場合は、申し出の内容に不備として継続審議扱いにしている

○ 今までの申し出は医療系に偏り過ぎ。工学系、政策系にも使ってほしい。

○ 第3者提供に応募する人はもう少し勉強してほしい。レセプトを触る実績を積んでから。
レセプトで何が出来て、何ができないのかが分かっていない申請者がいる。
一つのテーマを決めてそれに対する応募、はどうか？
練習用のDBを作って、これでうまくできたら申請出来るとか？
練習用ドキュメントはあるのですが…
検索を試してみたい
オンサイトセンターにするなら、練習は大事
抽出条件を作ったらそのまま申請書になるような、自分の作った条件が実行可能かどうかわかるような練習用DB

申出者と厚生労働省のコミュニケーション

○ 申し出者とのコミュニケーションをどのように充実させるか

有識者会議

(D) 申し出

○ これは本当にNDBでやる必要があるのかと思う申し出がある。有識者会議は要望に応えすぎ。
何をNDBでやるかを決め、目的別DBを作るための研究を累積
いちいち有識者会議で対応していたのでは収拾がつかなくなる

DataEffector

○ 3-4年運用して、DataEffectorのどこが妥当で、何処が足りなかったかわかるか
（富士電機）DataEffectorではCPU稼働がほぼ100%、I/Oがオーバーヘッドする

○ CPU稼働やオーバーヘッドは負荷テストでわかる筈。次の設計で手当てすべきことは？
（富士電機）2008年ではデータをためる仕様であり分析は含まれていなかった。
データ取り込みまでは最高速。(一ヵ月の要求を3日で実現)

○ I/Oがボトルネックなら、64並列を128にして処理をすれば倍に早くなるのか？
（富士電機）yes

○ I/O、CPU共にネックになっているのか？それともバカ高い/太いネットワークを入れれば解消するのか？
（富士電機）CPUは若干空いている。ランダムアクセスによるI/Oネックが発生しないように振り分けることが出来れば解消できるかも

○ DataEffector次versionでは解決できるのか
（富士電機）将来には共有、技術的に可能
技術の可能性じゃなくて、調達できるか
（富士電機）yes

○ NASを分けるとボトルネック解消できるか？
（富士電機）IとOを分散すれば。読まれるデータだけで作業ファイルに分ければ。
データをNAS1台で管理、ディスクアレイ2台にすれば同期はどうするか？　メンテナンスのコストが高くなる

○ 当面できることは、再匿名化,sortの問題、64並列が本当に最適なのかの検証

○ CPU100%利用しているがCPUの利用率を下げるとI/Oがぶつかってさらに遅くなる可能性あり。64を40と20くらいに分けて処理するのはどうか

H/W構成

○ 当面は中間生成物用NASを増設することで良いかも
2Tくらいの安いNASを追加で入れてみては？
上等なのは必要。でも、スループットを上げるために安いの入れる

○ 現状の64分割が足かせになっているのでは？
折衷策として、データ集積はローエンドにして(安くする)、高性能のマシンは解析系にシフトさせ、アドホックな要求にこたえられる運用にする

○ 機械の話では、現状13TB/3年。H26は、ペタ単位かテラ単位かでデザインが変わってくる

○ データの二重補完。レプリケートしておいて、クラウドサービスへ移行。

取り込みデータ

○ 取り込んだデータをそのまま格納する仕様か？
（富士電機）CSV形式。記録条件仕様そのままではなく、各ファイル種別でテーブルを分けてある。

○ 第3者利用、不定期集計が生じてきた現段階で、そのまま格納する仕様は良かったか？
（富士電機）MNテーブルに通し番号、実際はファイル種別ごとに格納。これが最速だった

データと使用頻度

○ HOT,COLDのデータ区別する。使わないデータは遅いところへ格納する(COLD)のだが、今の運用はすべてHOT？
（厚労省）全て並列にHOTでお願いしている

○ 月次バッチ処理をする時、過去データが入っている必要はないだろう
月次処理を独立させるのはどうか。64並列のうちいくつかを月次処理専用にして、他をNDB抽出用にする
（富士電機）可能

○ リサーチの為に最適でないデータ構成をしている。何かの目的別データモデルに対応したSQLを研究者に作ってもらうのが良いかと
（富士電機）格納データは再変換可能。

基本データセット

○ 基本データセットは、最低限のデータを数種類(何を入れるかは考える)、ランダムサンプリングした個人を経年で追えるようにしたい。

素DB

○ あるべきNDBの図で、素DBは本当にあるがまま、各テーブルに分けることもせず、記録条件仕様のままため込む
（厚労省）そのままなら媒体を保管しておくだけで良いのでは？
COLDくらいで保管する。凍結はだめ

○ 次期システムでは、素DBの段階ですぐさまデータを検証し、NGならば(国保連や支払基金から)再提出させる仕組みを
検証対象は個人ではなく保険者。ある月のある保険者の人数が多いとか。何をチェック対象とするかは、検討会などで決めていく。
そのチェックは国保連にやってもらっては？
今のフローでは難しい
国保連の業務の流れとして、決算システムさえよければ良いという考え。電レセは修正しない。

○ 欠損部の正規化をするなら、素DBは匿名化を行わず、欠損補填時に匿名化を。
素DBではデータはそのまま持っていて、厚労省で匿名化を

目的別DB

○ 目的別DB,どのような目的か？

○ 例えば、社会診療行為別調査や最適化計画に使ったデータセットを目的別DBとする

○ DBというかデータセットを、SPSSとか特定のソフトしか使えない人も使えるように、正規化しきっていない形で提供(入院、外来、調剤別など)

○ なるべくデータはrawレベルで出すのが重要。いろんな研究者がいろんな方向から自由にデータを使って研究できるように。
用語の統一が大事。

○ 研究者が自由に使えるCPUとデータを柔軟に組み合わせる。一方、月次処理はCPUとデータがセットになっていると良い。
HOT,COLDに分けて、3分で使えるデータ、1日、1ヵ月で使えるデータに分ける。

○ DPCの様式1が素DBに入らないか？
9/5の会議で様式1の個人情報とメリットの検討をしている。第3者提供はともかく、適正化計画に使える

○ EFファイルのように病院から支払者へ提出してもらうのどうか？
請求時にEFファイル化できないか？
EFファイルは活用しやすいし、病院のDB化の苦労も減る
（厚労省）長期的にはアリ。ただし、国保の決算システムまで変更するとした場合、膨大な…
レセプト様式を変えるのはアリ
転換時は費用が掛かっても、転換後は安くなるのでは？

目的別DB：オンサイト

○ オンサイトで使いやすいDBを。大規模だけNDBで個票抽出する。
（富士電機）目的別DBのいくつかをオンサイトセンターに。SQL、GUIで条件式を作る。
使う人の目的を絞り込む

○ 米国のMedPer方式が良い。研究者が個票を持ち、拡散するのは困る。集計データだけ(手軽に)持ち帰ってもらう。

目的別DB：定期集計データの活用

○ 定期集計の項目は、多くの臨床家にとって魅力的。もっと種類を多く、公表されれば、個票がなくても事足りる場合がある。
各学会が症例の登録を行っているが捕捉率が分からない。例えば、Kコード別都道府県別レセプト枚数が分かると役に立つ
ICD10、処理件数だけでも出して公表してみては
（厚労省）定期集計は公表してない、精査してない。活用されていない。
変な数字でないか、ちゃんと見て拡充する。データの精度管理大事。

○ どの統計量をopenにするかは、モノによっては個人情報がばれるリスクあり。
学会単位にリクエストしてみたらどうか

○ OECD ヘルスデータ、歯抜けになってる日本データを出すにもNDBは使える

マスター

○ 二次医療圏マスター等を研究者が用意するのではなく、あらかじめ用意されていると便利。研究者側も使えるようにする

(G) データセット

マスターメンテナンスが重要。データの活用可能性を上げる。
トータルデザインを議論すべき。
NDBに使用するマスターを整備すれば使いやすくなる。

(A) 運用

- □ DBは各作業で必ず富士電機のスタッフしか触れないのか？
  （厚労省）厚労省判断で制限している。データは個人情報なので外には出さない
  H26の再入札、入れ替えでどんなシステムにするか？を考える
- □ 申出書が申請されるたびに、富士電機に申請書を流せば、早く抽出にかかれる。
  しかし厚労省では限界がある。第三者が窓口代行をするか
- □ データを解釈した上で成長モデルを作る必要がある。
- □ reasonableなサブデータセットを作って小さくしてやってみる。こんなのが欲しいと決めてからベンダーへ。
  長期間とらないと分からないが、一旦、走り出したら早い筈。

(B) 匿名化、ハッシュ

- □ 紐付成功を増やすには？　共通番号はいつになるかわからない

(D) 申し出

- □ (文科省？の持ってる)スパコン使う人はSQL文をつけて申請することになっている
- □ 審査はアカデミックに依頼するなど、有識者会議だけでなく、外部の力を借りるのものあり。

(E) HW／SW

- □ DataEffectorの置き換えの必要性。オラクルだと研究者がSQLを書いて抽出条件を設定できる。
  トータルコストを下げるためのソリューションは？
  （富士電機）2008年の調達時点ではいったんのデータ格納システムとしてDataEffectorが良かった
  目的が決まれば選択肢が出せる
- □ HWのコストダウン、スピードアップ、しかしストレージが目いっぱい。
  昔のものは捨てなくないし、再利用もしたいの要求がある。

(H)その他

- □ どれくらいのトラフィックをどれだけで処理するか
  メンテナンスが重要
  クラウド利用すればどうか？

(G) データセット

- □ 現状の〇〇原本というDBは受け取ったままのデータか？
  （富士電機）yes
- □ 素DBからの加工は今のままで良いか？
  （富士電機）欠損値の補完は今でもやっている
- □ DPCの患者ID、マッチングを前提とした項目を検討したい

4. レセプト情報・特定健診等情報データベースの将来スキーム平成24年9月24日の合同研究会を元に考えられる今後の在り方

レセプト情報・特定健診等情報データベースの将来スキーム

平成 24 年 9 月 24 日の合同研究会を元に考えられる今後の在り方

(直近から平成 26 年以降に向けて)

1. 「中期的なアウトプットを具体的にデザインすることで、効率的なシステム設計を可能にする」
   - NDB の特徴を生かせる定期集計(月次、年次、社会医療診療行為別調査等を含む)、政策上の要求事項、想定される非定期集計、研究領域を明確にして、それに適した新しいシステムの設計を、現行システムにとらわれることなく行う。

2. 「ベンダーに依存しないシステムを構築する」
   - データ処理の手順やプログラムを NDB の運営組織に引き継ぎ、また蓄積できるような仕組みづくりを行う。
   - データ処理過程の全体の仕様において、特定のベンダーによるブラックボックス化を避けるようなオープンシステムを構築する。

3. 「定型、非定型を分けて、同時処理を可能にし、データ処理を効率的に行う」
   - 定型(月次処理)、非定型(NDB)に適したシステムをそれぞれ独立に構築し、並列処理を行うことで効率化を図る。
   - 例えば、定期処理は NoSQL、非定型処理は RDB を用いる。

4. 「研究に活用可能な定期集計データを集めたデータテーブルを充実、公開する」
   - 全件をカバーする NDB の特徴として定期集計データだけでも研究に資するものがあり、早急な公開が望まれる。
   - どのような集計データが有用であるかは、学会等の意見も聞き、検討していく必要がある。

(実現には予算を伴う NDB 周辺重要事項)

5. 「NDB 運用を担当する組織を常設する」
   - データの将来活用計画、保守運用、利用申請への対応を厚労省の少人数の担当者で行うことは限界があるため、今後いっそうの活用を進める上では、常設の組織が NDB の運用を担当することが望まれる。

6. 「各種マスターを維持管理する組織を持つ」
   - 必要とされるマスターテーブルは多数あり、例えば診療報酬改定等で随時更新されるものや、市町村合併等により変更されるものもある。
   - マスターのバージョン管理(更新および追加)に責任を持ってあたる専門組織が必要である。
   - 研究対象に即して最適なバージョンのマスター類が適用されることで、データ解析の精度が上がることが期待できる。

7. 「データ形式は研究に適した DPC E/F ファイル形式で構築する」
   - 既に DPC 参加病院では作成されているファイル形式であり、E/F ファイルを厚生労働省に提出している。
   - DPC に参加していない医療機関においては、レセコンにすでに E/F ファイルを作成する機能はあり、その機能を追加導入することで、ファイル作成が可能になる。
   - 各審査支払機関において一時的なシステム変更コストを要する一方、E/F ファイル形式は審査においても一層の効率化に資する可能性が大きい。

8. 「利用者の教育・研修のための組織を常設する」
   - 申請者へのサポート、オンサイトセンターの運営、人材育成等を担う組織を常設する。

2012年9月24日の研究会で検討されたレセプト情報・特定健診等情報データベースの現状の問題点と改善策

	現状	改善策	
		直近	平成26年にむけて
①H／W S／W	定期集計作業中は、CPUをほぼ100%使用するため、他の処理の同時並行作業が出来ず、作業効率が良くない。 NAS経由の集計及び抽出によるI/Oボトルネックが発生している。NDBでは64並列で処理を行っており、作業ファイル、抽出ファイル等、NASに同時書き込みを行う関係上、I/Oのボトルネックが発生する。抽出したファイル等は全てNASを経由するというシステム構成になっているため、回避できない。ただ、集計をするとき等は、各ブレードのサーバー内のほうで集計、マージを行って極力ファイルを小さくすることでボトルネックを最小限に抑えている。	•64並列を倍にして128並列にするとレセプトデータ取り込み作業時間は早くなると思われる。 •64並列を、例えば、抽出処理に40とデータ解析に20くらいに分け、別系統で処理するのはどうか。 •ネットワーク機能を補強する。 •NASをミラーリングすることによって取り込みと抽出の同時並行作業が可能になる。(但し、同期、メンテナンスコストが高くなる) •中間生成物用NASを導入する。(2TBくらいの安いNASを追加) •データ集積はローエンドにして(安くする)、高機能のマシンは解析系にシフトさせる。	•(ベンダー意見)現行ソフトウェアの次バージョンで、CPU100%使用の問題が解決可能である。具体的に調達であるとのこと。 •将来、ペタ単位かテラ単位か、どちらを想定するかによってデザインが変わる。(現在は13TB/3年)
②データ格納方法	すぐに使うデータも、たまにしか使わないデータも並列にHOTで運用している。	•データは複製しておいてクラウドサービスへ移行する。 •月次処理を独立させる。64並列のいくつかを月次など定期的な処理専用にし、他を非定期の処理、NDB申し出対応用にする。 •データをHOT,COLDに分類し、例えば3分で使えるデータ、1日、1ケ月で使えるデータに分ける。	
③データセット	データ抽出、分析、活用に適していないデータ構造になっている。	•抽出、分析を行う者から見て容易な中間モデルの構築を行う。	•国保連、支払基金からのデータ取り込み、蓄積先として、レセプトデータ集積用一次DB(csv)を作成する。 •レセプトデータ集積用一次DBには記録条件仕様のまま蓄積する。 •DPC E/Fファイルの形式で蓄積する。 •DPC様式1も入れる。(9/5の有識者会議で検討している) •DPC 様式1に準ずるものを外来にも導入する。 •匿名化を行わない。
			•目的別データベースを作成する。 •例えば、社会診療行為別調査や最適化計画に利用したデータセットを目的別DBとする。 •入院、外来、調剤別などに分ける。 •なるべくrawレベルで提供する。 •ランダムサンプリングした個人を経年で追えるようにする。
			•データを制限した使いやすいDBを構築し、オンサイトで研究者がデータを抽出できるようにする。大規模な研究だけNDBで個票抽出を行う。 •オンサイトでは、研究者は集計データだけ持ち帰るようにする。 •研究機関等にオンサイトで研究者がデータを抽出できるオンサイトセンターを構築。NDBの運営、人材の育成、マスター管理も同時に行う。

	定期集計データが公表されていない。	•定期集計データの活用を検討する。 •データを精査して、活用されるようにする。データの精度管理が大切になる。	

④匿名化とレセプトデータ	ハッシュ値の正確さに不安がある。同一人物と思われる人がID1,ID2で繋がらないことがある。	•データ収集時には匿名化せず、データを外に出す時に初めてデータセンターで匿名化を行う。 •レセプトデータの段階で同一人物のレセプトを識別できるようにする。	•どれくらい正しくあるべきと考えるか、現実的な期待はどの程度か共通の認識を持つ。
	IDだけでカウントすると1億4千万人くらいになる。(日本の人口より多い)		
	ハッシュ値の正確さが完全でないために、既存統計より精度がおちている懸念がある。		
	今は医科だけだが、今後は介護データともつなげたい。		
	レセプトデータ作成において、市町村国保でデータチェックの厳しさが違う。	•データ受付時にチェックを行う。そのために、チェック項目をリストアップする。(レセプト単位、保険者単位) •何をチェック対象とするか、検討会等で決める。	•ID作成の元になる氏名が漢字であるため、表記ブレが起こっている可能性がある。氏名をカタカナにする。 •レセプトデータの段階で個人を同定し、つなげられるようにレセプトデータの精度を上げていく。 •医療機関の入力の段階から定期的にデータを検証する機会を作る。
	国保連は提出済のレセプトデータを次々と消していくので、問題が発覚した時点によっては、国保連まで戻って検証できない。		
	被保険者の保険が変わり、氏名が変わった場合、今のハッシュ値のつけ方では、追跡できない。(例、結婚し、退職し、転居した場合)		•保険者が被保険者の異動を報告する仕組みを作る。
	日本語のコメント部分が全部削除されている。	•コメント部分を残す。	
	レセプトに退院日、再入院日がない。		•DPC 様式1には、退院年月日、前回退院年月日があるので、DPC 様式1+E/Fファイルでデータを蓄積する。
	レセプト傷病名のつけ方がいい加減。		•DPC 様式1に準ずるものを外来にも導入する。 •(レセプトデータを分析した上で)重みづけロジックの実装を検討する。
	患者数が少ない症例はいい加減なデータが入りにくい。		
	記録条件仕様に返戻フラグを表す項目がないので、返戻フラグの有無が分からない。	•返戻フラグを保持できるデータ構造にする。 •返戻がかかっていることが分かるフラグを追加する。(データの取捨選択は研究者が行う)	
	返戻レセプトは最長で2年かかるので一次審査データを使うのは妥当だと考えられる。		
	(2010 年からは返戻レセプトを取り込まないようにしている。)		

⑤利用申し出への対応	抽出条件からデータ量の予測がつかない。		•研究機関等にオンサイトセンターを構築する。オンサイトで研究者が自らデータ抽出を行い、抽出条件を検討する。
	データ抽出条件に書かれた、あいまいな仕様の確認に時間がかかっている。 申出者のレセプトデータに対する理解が不足している。(できることとできないことが分かっていない)	•データ抽出用 SQL 文を提出するようにする。SQL 文がそのまま抽出に利用できるので、オペレータの介在時間が減る。しかし、今のレセプトデータの仕様では研究者が SQL 文を作るのは難しい。 •練習用 DB で抽出が出来るようになったら申請できるようにする。 •レセプトデータ解析を行った経験を申し出資格にする。	
	有識者会議がNDBで行う必要のない研究にも応えている。 今後も増えるだろう申し出すべてに応えられない。	•NDB を用いるにふさわしい研究の要件をより具体的に絞る。 •例えばテーマを決めて、テーマに関して公募する。	•大規模な研究だけ NDB で個票抽出を行い、その他の研究は研究者が自らオンサイトセンターを利用し、データ抽出、集計を行う。

⑥運用	研究者とベンダーで直接コミュニケーションの場がなく、意識の乖離が大きい。 研究者、ベンダー、厚生労働省(有識者会議)で言葉が通じていないことがある。	•用語の統一を図る。 •ベンダー、研究者が共用できるマスターを整備する。 •(試験的に作成している)依頼用テンプレートを用いることによって、効率が上がるかを確認する。	•研究機関等に研究の申し出の窓口となる組織を常設する。
	NDB 申し出対応でデータ抽出時に、名寄せ、ソート、採番しなおしの過程に時間がかかりすぎている。 (抽出自体は 10 時間ぐらいでできるが、通番の1、2を再匿名化するルールになっているため、再匿名化のところで、1回全部集めて、ソートかけて振り直す処理に時間がかかり、トータル 200 時間くらいかかってしまう。)	•ハッシュのつけ方や、ソートをしなくても良いようにアルゴリズムを改善すれば処理は早くなると思われる。 •匿名化の項目※が多いために時間がかかると考えられるので、匿名化の項目を減らす、またはなくす。(※管理番号、通番 1,2、保険者番号、レセプトレコード番号等)	
	研究者へ課されたセキュリティ要件が部分的に過剰に厳しい。	•抽出データは匿名化されていることを踏まえ、セキュリティ要件の一部を緩和する。(例えば、加工・集計したデータであれば、申し出者以外が在席する研究室等での利用を可能にする)	

5. レセプト情報・特定健診等情報データベースの運用について（プロセス・マップ）

レセプト情報・特定健診等情報データベースの運用について

000	ベンダーまで			ベンダー：データ取り込み 001								ベンダー：集計・抽出 001					
NO.				1, 2	3, 4, 5, 6,	7, 8, 9, 10, 11	12, 13	14, 15,	16, 17	18, 19	20, 21	22	1, 2, 3, 4, 5, 6	7, 8, 9, 10, 11, 12, 13	14	15, 16	17, 18, 19	20, 21, 22, 23	24, 25, 26, 27, 28
作業担当	審査支払機関	審査支払機関→厚労省	厚労省→データC	オペC	データC	オペC	オペC	オペC	オペC	オペC	オペC	オペC	厚労省・オペC	厚労省・オペC	オペC	オペC	オペC	オペC	厚労省・オペC
作業	匿名化	媒体送付	媒体送付	媒体装填依頼	媒体装填	媒体展開	NASへデータコピー	ファイル分割	名寄せ処理実施	定期集計	レセプト数集計	取込結果報告	集計依頼書の送付、登録	抽出条件の確認	定義体の作成	テストデータ作成	テストデータ集計	本集計	データの~~複写~~、~~暗~~号化、媒体書き込み、送付
所要時間										~~約半日~~→1日		（1-22で約3.5日）							
追加説明	・各審査支払機関、各国保連（一部は健保組合が直接）において実施	・前月審査分が、当月月末に届く。	・厚労省からはできるだけ早期にデータCに送付する。	・厚労省→オペC→データCと伝書	・媒体の装填	・解凍 ・展開されたことを確認	・NASへデータを移し替える。	・レセプトデータを64ファイルにできるだけ均等に格納する。 ※1レセプト単位で分割	・データに年齢を付与し、二度目のハッシュ値を附与する。 また、集計上必要な通番（レセプト単位で一意となるキー）とレセプト内のレコード順も項目として追加	・60-70項目にわたる集計を毎月行っている。	・毎月の単純集計を行いメールで通知	・現在は、このあとに調剤メディアス向けデータ抽出・提供（1日）が行われている	・抽出条件を提出する。 ・抽出条件を厚労省との間で確認する。		・抽出条件を~~データイフェクター~~ DataEffectorで作業できる記述に改める。	・定義体の作動を確認するために最適なミニデータを作成する。	・作成した定義体をミニデータにかけて、整合性を確かめる。	・抽出条件を全データに適用して抽出する。	
今後の課題	・ハッシュ値の精度向上 ・二次審査データの差し替え ・重複してカウントされているデータの処理									・あり方の見直し		・件数以外、どういった箇所の内容を確認すれば精度の確保につなげられるか。							

100

(A) 運用
(B) 匿名化、ハッシュ(個人特定)
(C) レセプトデータ
(D) 申し出
(E) DataEffector　HW
(F) データ格納方法
(G) データセット

レセプト情報・特定健診等情報データベースの運用について

000	ベンダーまで			ベンダー：データ取り込み 001								ベンダー：集計・抽出 001					
NO.				1, 2	3, 4, 5, 6,	7, 8, 9, 10, 11	12, 13	14, 15,	16, 17	18, 19	20, 21	22	1, 2, 3, 4, 5, 6	7, 8, 9, 10, 11, 12, 13	14	15, 16	17, 18, 19	20, 21, 22, 23	24, 25, 26, 27, 28
作業担当	審査支払機関	審査支払機関→厚労省	厚労省→データC	オペC	データC	オペC	オペC	オペC	オペC	オペC	オペC	オペC	厚労省・オペC	厚労省・オペC	オペC	オペC	オペC	オペC	厚労省・オペC
作業	匿名化	媒体送付	媒体送付	媒体装填依頼	媒体装填	媒体展開	NASへデータコピー	ファイル分割	名寄せ処理実施	定期集計	レセプト数集計	取込結果報告	集計依頼書の送付、登録	抽出条件の確認	定義体の作成	テストデータ作成	テストデータ集計	本集計	データの~~複写~~暗号化、媒体書き込み、送付
所要時間										~~約半日~~一日		(1-22で約3.5日)							
追加説明	・各審査支払機関、各国保連（一部は健保組合が直接）において実施	・前月審査分が、当月月末に届く。	・厚労省からはできるだけ早期にデータCに送付する。	・厚労省→オペC→データCと伝言	・媒体の装填	・解凍 ・展開されたことを確認	・NASへデータを移し替える。	・レセプトデータを64ファイルにできるだけ均等に格納する。 ※１レセプト単位で分割	・データに年齢を付与し、二度目のハッシュ値を附与する。 また、集計上必要な通番（レセプト単位で一意となるキー）とレセプト内のレコード順も項目として追加	・60-70項目にわたる集計を毎月行っている。	・毎月の単純集計を行いメールで通知	・現在は、このあとに調剤メディアス向けデータ抽出・提供（1日）が行われている	・抽出条件を提出する。 ・抽出条件を厚労省との間で確認する。		・抽出条件を~~データイフェクター~~DataEffectorで作業できる記述に改める。	・定義体の作動を確認するために最適なミニデータを作成する。	・作成した定義体をミニデータにかけて、整合性を確かめる。	・抽出条件を全データに適用して抽出する。	
参考となる事例	・国保連由来のデータの一部でデータ取り込み時のエラーがあった。					・データの一部で展開未了のまま終了したことあり。 002							・申出によっては、内容の確認で厚労省との間で何度もやりとりあり。		・模擬申出の1例：約３カ月→サンプリングデータセットでは約１カ月に短縮 010			・承諾申出の１例：約200時間 ・データセンターに厚労省専属のスタッフはいないため、オーダーの受付が少し遅れることあり。	・DVDとHDDで複写時間が10倍ほど差がつく。 ※書き込み時間以外でも、DVDでは、IDC作業員が媒体を一枚毎に装填、オペCのメンバーが書きこむという作業が発生する為、この時間もロスとなる（HDDは接続の
遅延が発生する可能性	・基金、国保中央会からデータ提供が遅れたことはない。 ・一部健保組合からのデータ提供が遅延することがある。											・取り込み作業と抽出作業を同時に行う事ができない。 003	・抽出条件の頻回の確認		・オペレーションセンターの人員が限られている。 004 ・レセプト構造に起因する課題がある。 005			・集計表は比較的すぐ出せるが、個票形式で提供する場合には、出力ファイル量が多くなる為、時間がかかる。	
業務簡略化の可能性											・定期集計との違いは？		・数ヶ月分の抽出まで極力前もって提 006 ・依頼書の書式をオペCが把握しやすい形式に改める		・過去の抽出条件の分析を行い、共通の抽出ロジックについてパターン化する。			・IDの振り直し（医療機関コード、保険者番号、ハッシュ１、２）の改善 ※現在はシステム内コードをそのまま出さない事としている為、別コードに振り替えているが、処理時間が抽出レセプト数に依存する為、大量に抽出される場合、この処理に時間がかか	・申出者にHDDの提供を義務化
質問事項その他															・どんな作業か？（判りやすい具体例は？） ・実際にどの程度時間がかかるか？ ・テストデータは毎回作っているのか？	・実際にどの程度時間がかかるか？			

レセプト情報・特定健診等情報データベースの運用について

000	ベンダーまで			ベンダー：データ取り込み 001									ベンダー：集計・抽出 001						
今後の課題	・ハッシュ値の精度向上 ・二次審査データの差し替え ・重複してカウントされているデータの処理									・あり方の見直し		・件数以外、どういった箇所の内容を確認すれば精度の確保につなげられるか。							
																			100

(A) 運用

(B) 匿名化、ハッシュ(個人特定)

(C) レセプトデータ

(D) 申し出

(E) DataEffector　HW

(F) データ格納方法

(G) データセット