教育・研究のためのデータ連携ワークショップ

医療・健康データ中心科学の実際
データクレンジングとマイニング

倉本 秋・片岡 浩巳
高知大学医学部

医療･健康データの発生元と統合DWH構築の課題

施設間誤差
測定方法間誤差
検体輸送時の誤差
検体変質の誤差

健康分野

医療分野

健診センター

複数の外部分析施設

病院、診療所

個人識別は数年後

特定機能健診

特定保健指導

まだ大部分のシステム間の
データ交換は、紙ベースの運用

すでにオンライン収集
手順が確立され実施され
れている分野
しかし、これらは連携さ
れていない、個別のプ
ロジェクト

院内感染
サーベランス

DPC関連データ

診療データ

施設間誤差
測定方法間誤差

個人識別ばらばら

統合は、これから

共通の電子カルテ

家庭

グループ毎に分散された
データベース

個人（モバイル血圧計など）

統合したデータウエアハウス

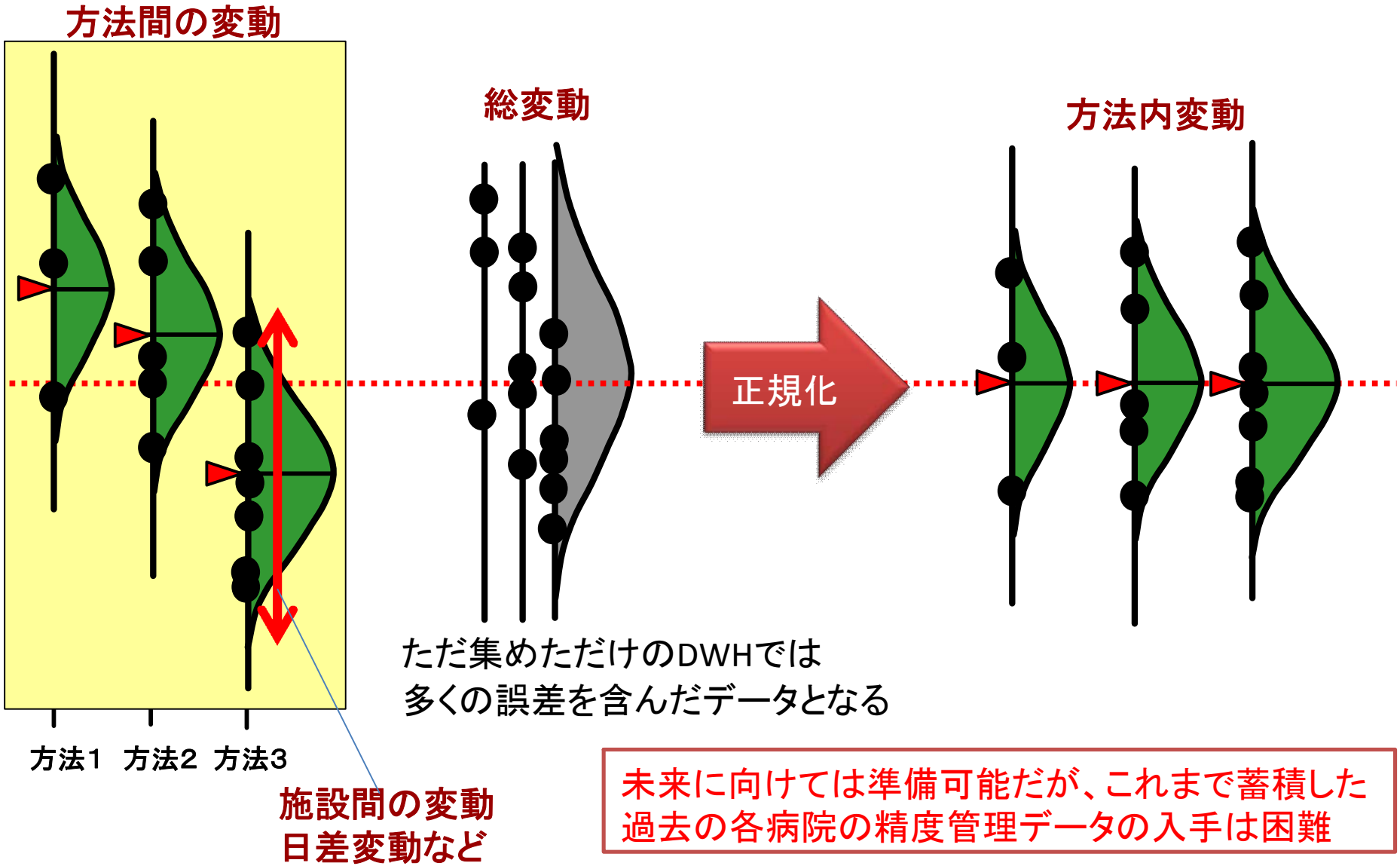
多くのバイアスを含むデータ

プロトコルの標準化は完了し
ているがデータの正規化の
課題は多い

単にデータを集めただけの
DWHではマイニングは不能

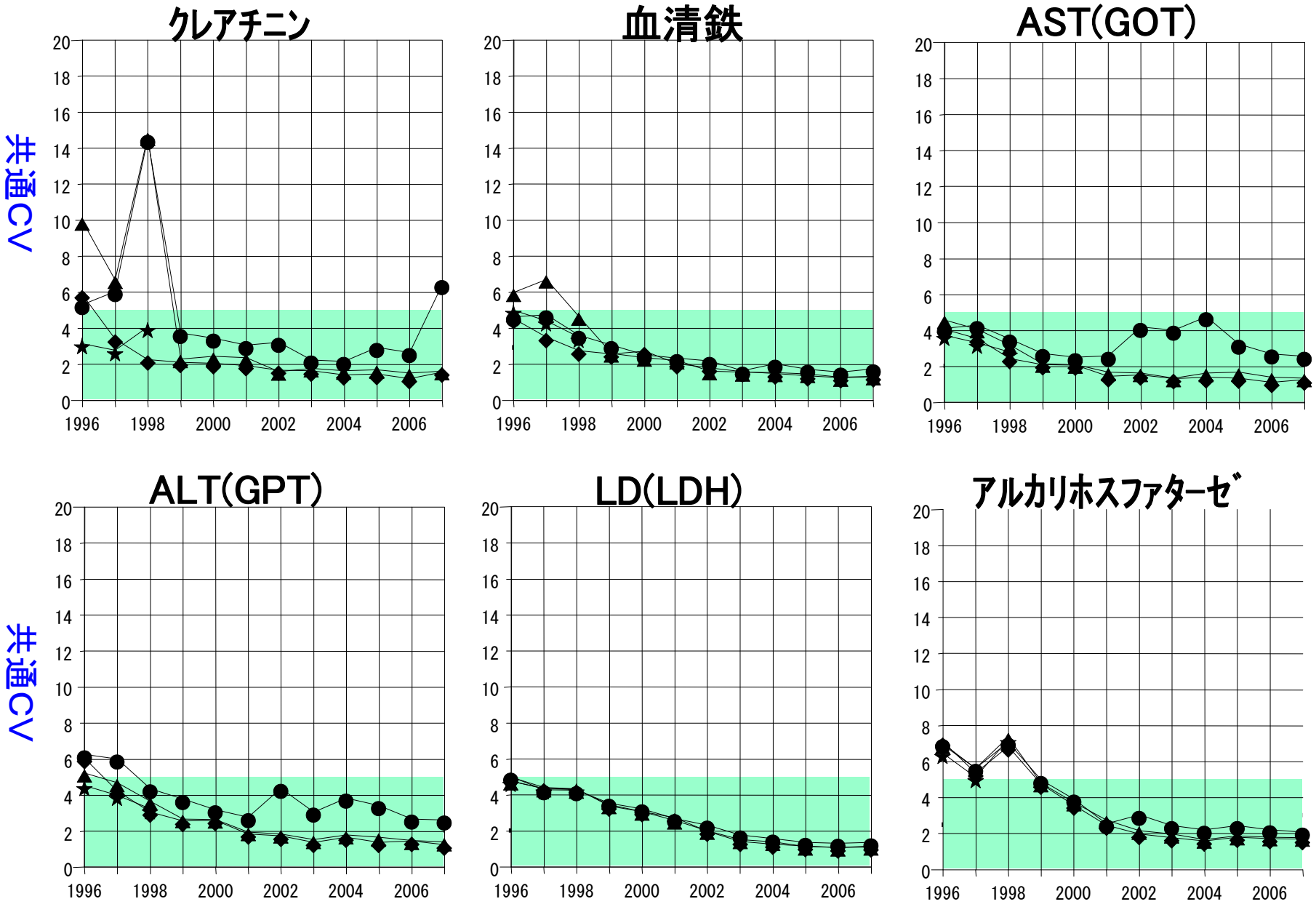
方法間、施設間変動によるバイアスのクレンジング

外部精度管理（同一の試料をすべての分析施設で測定した場合）



未来に向けては準備可能だが、これまで蓄積した過去の各病院の精度管理データの入手は困難

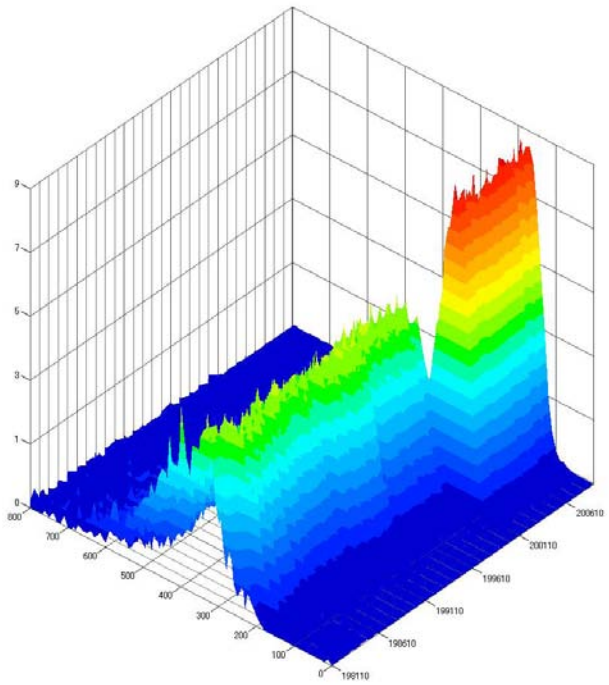
全国・外部精度管理の実態　過去と今

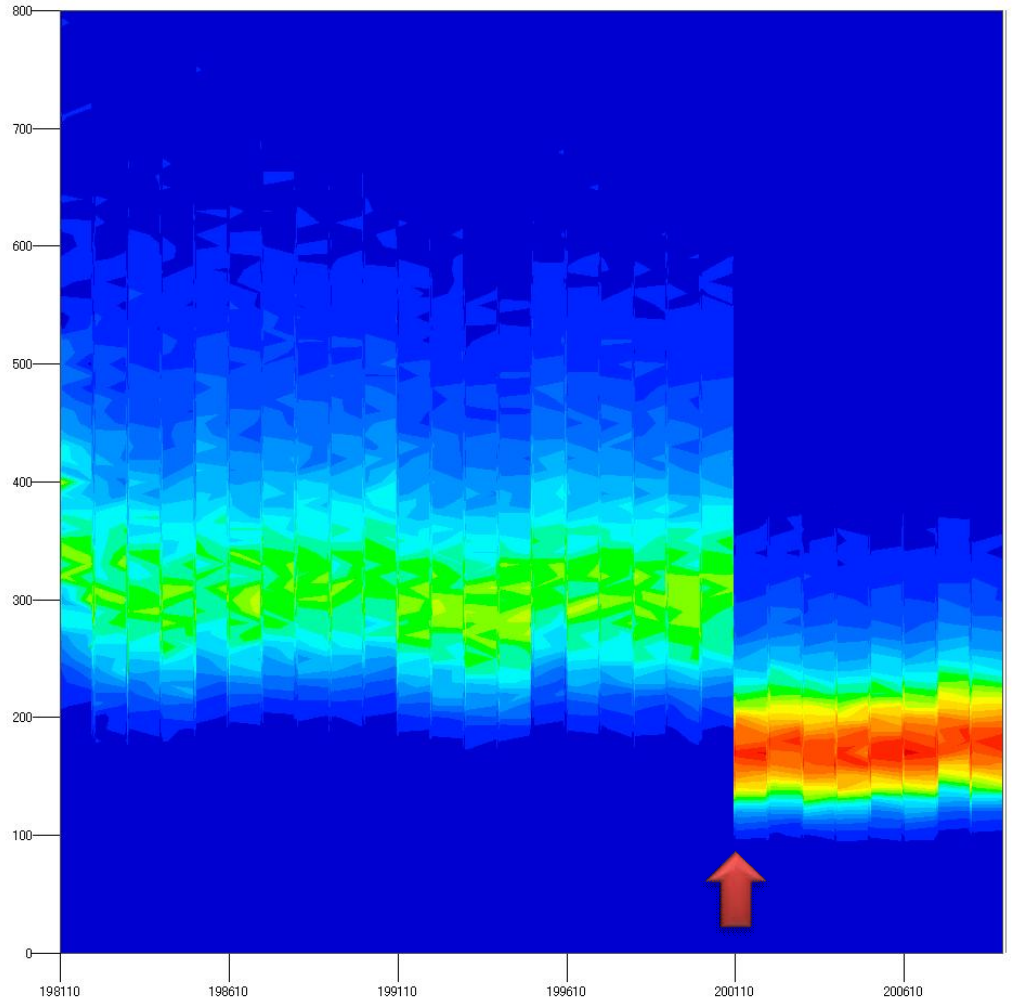


日本医師会 外部精度管理報告書より引用

高知大学28年間の臨床検査値の分布

乳酸脱水素酵素(LD)の28年間にわたる集団分布（年月毎に度数分布を作成）

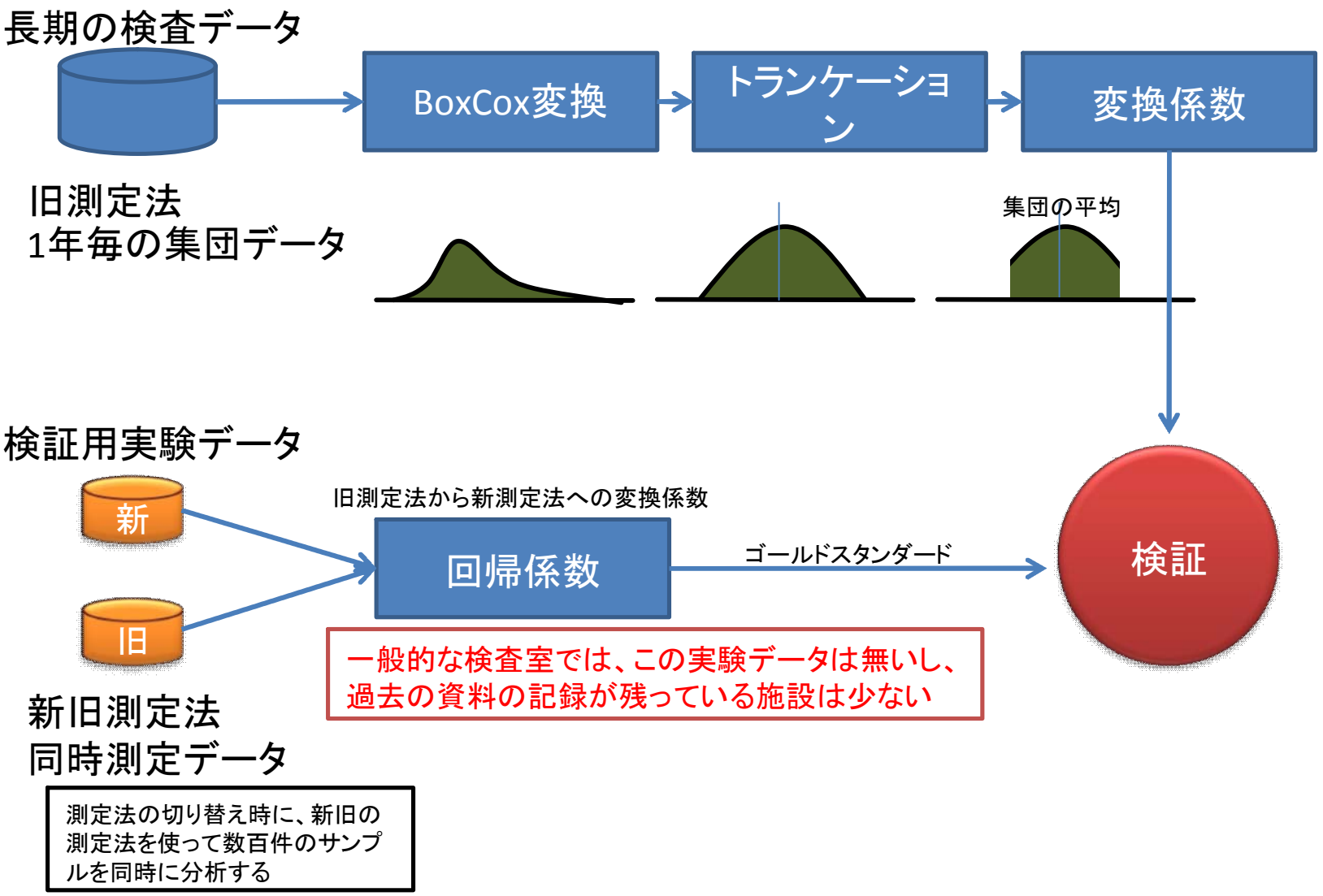




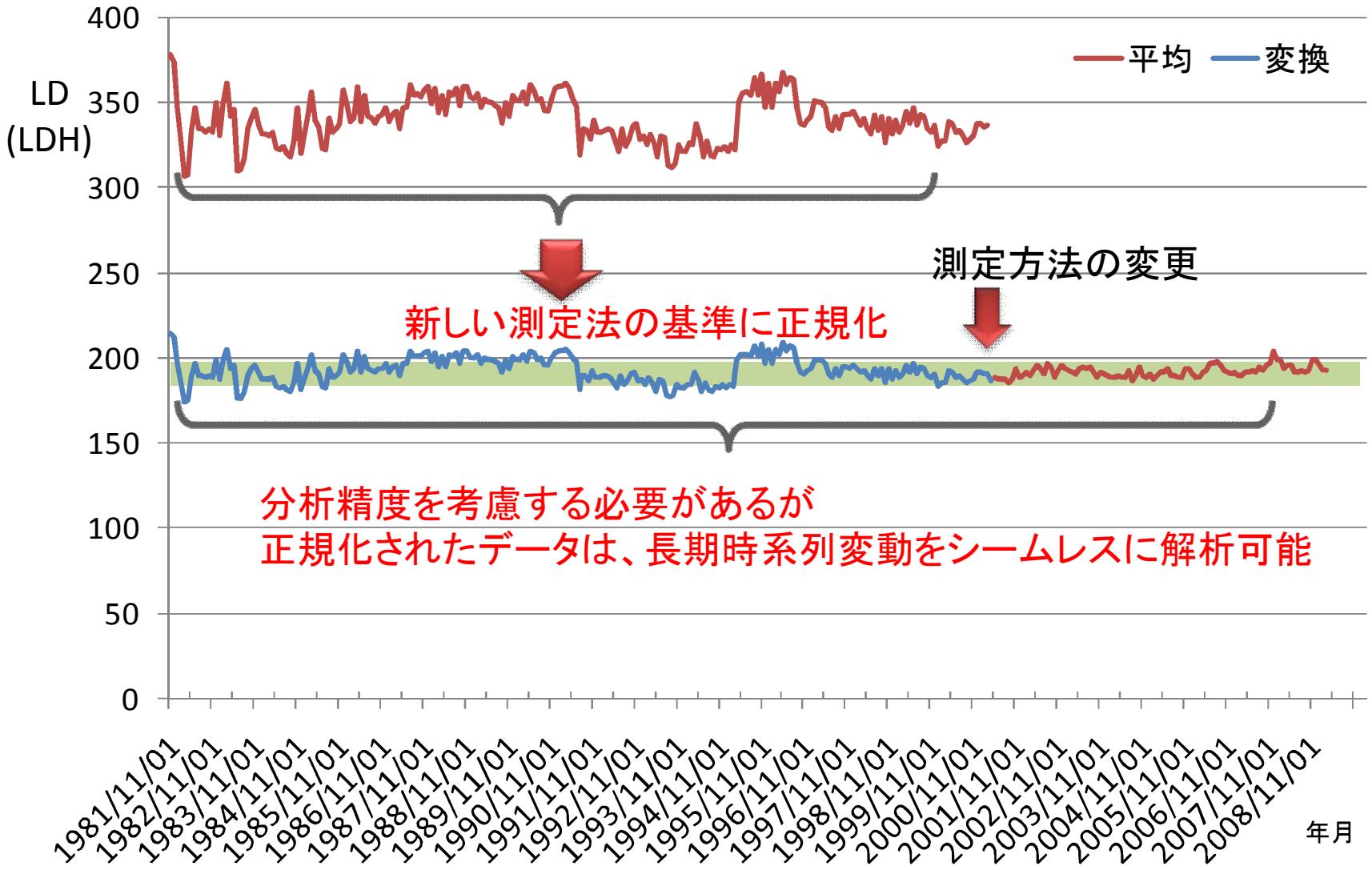
測定方法の変更

検査データの正規化技術の開発

目的：蓄積された膨大なデータの分布から、変換のための係数を求める



正規化処理結果


400
350
300
250
200
150
100
50
0
LD
(LDH)
平均
変換
測定方法の変更
新しい測定法の基準に正規化
分析精度を考慮する必要があるが
正規化されたデータは、長期時系列変動をシームレスに解析可能
1981/11/01
1982/11/01
1983/11/01
1984/11/01
1985/11/01
1986/11/01
1987/11/01
1988/11/01
1989/11/01
1990/11/01
1991/11/01
1992/11/01
1993/11/01
1994/11/01
1995/11/01
1996/11/01
1997/11/01
1998/11/01
1999/11/01
2000/11/01
2001/11/01
2002/11/01
2003/11/01
2004/11/01
2005/11/01
2006/11/01
2007/11/01
2008/11/01
年月

まとめ

- 臨床検査データのクレンジング法の開発
  - 精度管理情報がなくても補正可能な手法
  - 蓄積されたデータの分布形を正規化し、その平均値を指標として変換
  - 実用レベルの成績が得られた

- 全国規模のDWH構築に向けての課題
  - 測定法、分析施設などの詳細情報を検査値と同時に収集、管理する必要がある