ある医療データベースの統合

神奈川県立こども医療センター
血液･再生医療科
田渕　健

2008 年 12 月, 三菱東京 UFJ 銀行のシステム統合が完了したとの報道があった. 総費用 3300 億円, 開発工数 14 万人月に及ぶ巨大なプロジェクトであったという. 私もその銀行の利用者であったので, 銀行のシステムならば, 同じようなもののはずなのに, それを統合すると言うだけで, 何で, 1 日以上もシステムが止まらなければならないのかと思ったのが率直なところである. システムの統合がそんなに大変なことなのか, もっとエレガントな数学的な解法はないものなのか, 疑問に思ったのだが, 実は, 規模は全く異なるが, 私もある医療データベースの統合問題にかかわってきたので, 少しばかり述べさせていただきたいと思う.

私の仕事は臨床現場の小児科医である. 近年, 医療の分野でも細分化が進んで, 小児科の中でも, 主に小児血液腫瘍疾患の診療を担当している. 小児白血病, 小児がん, 貧血, 血友病といった病気である. この様な病気の中には,「造血幹細胞移植」という治療法が必要な場合がある. 造血幹細胞移植というと以前は骨髄移植に代表されたが, 最近では, 末梢血幹細胞移植や臍帯血移植が普及してきたため, 総称して呼ばれている. 造血幹細胞とは, 血液細胞のもととなる細胞のことである. 白血病やがんのように, 悪性度の高い細胞を取り除くための強力な治療を行うために, 適切に適合(HLA が一致ないしほとんど一致)している他人の造血幹細胞を移植したり, 固形がんでは, 予め自己の造血幹細胞を凍結保存しておいて, 強力な治療後に解凍して戻すという使い方もする. 再生不良性貧血というような自身の造血幹細胞が故障している場合には, 他人のうまく適合した造血幹細胞と取り替える必要がある. 造血幹細胞移植は, 日本では 1980 年代から積極的に行われるようになり, 今では, その治療が必要とされた場合には, 標準治療となりつつある.

しかし, 造血幹細胞移植は, 大変な治療である. 患者さんがその治療を受けるには, 相当な決心を必要とする. 移植直前の 1 週間位前からは, 多くの場合, 大量の抗癌剤や放射線治療を受けるため, それだけでも様々な臓器合併症が起こる. その上, 他人の造血幹細胞を用いる場合には, GVHD(移植片対宿主病)という一種の免疫反応が起こる. 免疫というのは, 簡単に言えば, 自己と非自己を認識するプロセスであるが, 移植された幹細胞から発生したリンパ球は, 他人の体には言った場合, 実は自分本来の体ではないことに気がつき, ここは, 自分の居場所ではないと思って騒ぎ出して, 移植された人の体に攻撃を仕掛けるのである. 移植しても, 原病が再発することもある. 対象の抗癌剤の影響で, 不妊になったり, 小さいこどもの場合には, 身長が伸びなかったりする.

移植の種類による生存率の比較や移植後の様々な合併症について, 完全に解明

されているわけではない．むしろ，未解決の問題が多い．それでも，造血幹細胞移植が標準治療となっているのは，他に救命可能な代替の治療法がないからである．このような未解決な問題を多く抱えている医療では，行われた治療の全てを 1 件 1 件把握して，問題点を徹底的に追求できるような情報が透明化されることが望ましい．しかし，医療情報のデータベース化は意外と遅れている．造血幹細胞移植医療は健康保険で認められたいわば「標準的治療」であるのなら，この様なデータベース化の仕事は，どこか公的機関がこのような仕事をしてくれるのかといえば，実は，そうではない．それどころか，移植に限らず，疾患登録でさえ，まだきちんとはなされていないのが実情であり，どのような疾患が何件ぐらい日本で発生しているのか，というようなことですら，正確な数字はわからないのである．情報というのは，生ものであるから，発生したら速やかにデータベースという形にしておかないと，消え去ってしまい形に残らない．つまり登録の母体がないとデータベース化されない．

日常臨床が多忙な中でそのようなボランティアをかってでようとする医師はそうはいない．小児科のように人手不足が深刻な場合にはなおさらである．一部自分興味だけでデータを集めて，独り占めにして自分の業績にしようとする輩はいるものの，自らの興味がなくなれば続けられないし，そう言うことが分かれば，各臨床現場の医師はデータを提供しようという気にはとてもならない．客観的に医療情報を管理する学問分野が必要であるが，日本ではそのような学問分野はなかなか認知されてこなかった．医療情報学，医療統計学(生物統計学)と呼ばれる分野である．こういう分野を学ぶ人たちは大学学部の数学科の知識は前提とされる．医療統計学は最近では医師の間にも少しずつ認知されるようになって，多少の誤用も散見されるが，曲がりなりにも使われるようになった．しかし，統計を行う前段階のデータベース自体がしっかりしていないと，データそのものの意味がぐらついてしまう．母集団を限れば様々なデータベースはある．しかし，それでは，偏ったデータの可能性がある．登録システムの仕事は，まずは，患者さんの知りたい過去の情報を提供する資料となる網羅的なデータを収集することである．このプロセスで，医療の透明化，情報公開を伴うことは言うまでもない．後方視的データ(実際に行われた医療情報の収集)であるから，臨床試験のような前方視的データよりもエビデンスとしては弱いが，網羅的に行えば，医学的研究にもきわめて有用であり，今後行うべき前方視的臨床試験の基礎データとなる．この仕事は，自らが **something new** を発見し，公表する仕事ではなく，むしろ，それを支援する仕事で，いかにも縁の下の仕事である．登録システムの大半の仕事はデータベースの基礎を作る仕事であった．

造血幹細胞移植登録事業について言えば，小児科領域では 1983 年から細々と続いてきたが，成人領域では 1991 年からのものしか網羅的には得られていない．1990 年代後半からは骨髄バンクや臍帯血バンクといったデータベースも形成されてきて，日本における造血幹細胞移植に関する登録システムは 4 つとなった．私は，この小児

科領域の造血幹細胞移植登録集計事業に関して，この 10 年ほど携わってきた．仕事の内容としては，造血幹細胞移植のデータベースを管理し，移植種類や疾患別の件数やその生存率を算出するという仕事である．

登録システムの分立によって，登録業務が煩雑になるという医療現場の負担に加えて，それぞれのデータの比較が，生物統計学的にバイアスが入ることがあげられる．後者は，患者さんにとって，正しい移植成績の比較の情報が得られないという不利益を得る．このため，このデータベースを統一しようということになった．4 年の歳月をかけ 2008 年末，ようやく日本における造血幹細胞移植データの一元管理が形をなすに至った．

データベースが異なれば，フィールドの定義が異なるため，それぞれのフィールドの対照表を作成すれば，あとは，プログラムするだけであろう，と最初は，もっと安易に考えていた．対照表が複雑なものになることは最初から予想していた．しかし，プログラミングで対処できない内容の存在が最も難物であることがわかった．それは，言葉で述べれば，1 件 1 件あたって「医学的判断を要する」対象と言うことになる．私は，金融機関のシステム統合がどういうものかは知らない．しかし，恐らく，このような難物はなかったのではないかと想像する．

医学的判断を回避するもっともカンタンな方法は，「その他」に分類することである．それならば，プログラマブルである．このように分けられた「その他」情報の多くは，あってもなくてもよいようなコメントであることも少なくない．しかし，内在する重要な医学的情報をそぎ落とす可能性がある．この「その他」に入ってしまうかもしれない情報から有意義な情報を拾い出して，実は見落とされていた情報を「その他」以外の部分からも拾い出すことがある．新たな entity の認識である．この段階で，新たなデータベースフィールドを定義する場合もある．新たな数学的な構造の発見に似た喜びがある．このほかに，フィールドの定義の予想を超えるような値やあいまいな値がしばしば存在した．外れ値や不適切な値と決めつけることが出来れば，よいのだけれども，データクリーニングの問題なのか，医学的には正しく，データベースの定義が不十分であるのか，を判断する必要も迫られた．

このようなデータマイニング的作業は，当面は，「医学的判断」として，ほとんど試行錯誤的な，気の遠くなるような作業の繰り返しである．私が担当したのは，小児科医であるので，小児科領域の造血幹細胞移植約 9000 件弱の移植データに関してであったが，プログラマブルな作業以外に 1 年近くの年月を要した．

登録システムが一元的でないと困る端的な例を示そう．生存率の比較を行うのに，登録システムのバイアスが存在することを触れたが，同じ対象でも，調査の時点が異なると，生存率が異なる可能性がある．例えば，白血病という病気では，骨髄バンクからの骨髄移植と臍帯血バンクからの臍帯血移植のどちらを選択したらよいかという問題が生じることがある．それぞれの登録システムで収集されたデータで比較可能であ

ろうかという問題が生じる. 生存率の推定は, ある指定された調査時点迄の間で, 特定出来る時点で亡くなっているのか(イベント発生), それとも特定出来る時点で生存していてそれ以降調査時点までの間の情報はないか(センサー), いずれかである. 調査時点後になれば, 当然, センサー群の中にイベント発生が現れうるし, あるいは, 生存時間が長くなるかいずれかである. 従って, 同じ対象でも異なる. A という登録システムの昨年調査の生存率と B という登録システムの本年調査の生存曲線を比較するのは, 不適切なのは, 直感的にもわかるだろう. しかし, 一元化以前の状況というのは, これに近い可能性があったのである. 調査時点の異なる 2 つの類似した調査対象を併せて解析することは, 統計ソフトウェア的には可能であっても, 結果の解釈には困難が伴う. これでは, 患者さんに還元する情報も混乱してしまう. 場合によっては, 別の結果を提示してしまうからである.

移植統計の話題は欧文の生物統計関係の雑誌には多く述べられているので, ここではこれ以上は述べない. 一元化された移植データベースは, 今後進化して行くには, 造血幹細胞移植治療が自己完結的な治療ではなく, 移植前後, 特に長期にわたって様々な経過を辿るため, 移植医療の真実の姿を知るには, 複雑な経過が盛り込まれたより質の高い移植データベースに発展させていくのが望ましく, まだまだ課題が山積している.