[**I.** **DATA MINING IN HEALCARE: CURRENT APPLICATIONS AND ISSUES** (By Ruben D. Canlas Jr., MSIT, MBA 5 August 2009) 2](#_Toc336589074)

[1. Trích dẫn: 2](#_Toc336589075)

[2. Giới thiệu và các nhân tố cơ bản: 2](#_Toc336589076)

[3. Phương pháp luận: 2](#_Toc336589077)

[4. Các nghiên cứu: 2](#_Toc336589078)

[5. Các vấn đề và thử thách: 6](#_Toc336589079)

[- 6](#_Toc336589080)

[6. Kết luận và khuyến nghị: 6](#_Toc336589081)

[**II.** **NEXT** 6](#_Toc336589082)

1. **DATA MINING IN HEALCARE: CURRENT APPLICATIONS AND ISSUES** (By Ruben D. Canlas Jr., MSIT, MBA 5 August 2009)
2. Trích dẫn:

* Với sự thành công của các ứng dụng được tích hợp Data mining trong các lĩnh vực rõ rệt như thương mại điện tử, marketing và bán lẻ đã cho ta thấy tính phổ biến của việc sử dụng những tri thức được lấy từ cơ sở dữ liệu ( Knowledge Discovery in Database – KDD) trong tất cả các lĩnh vực. Một trong các lĩnh vực mới nhất vừa được áp dụng Data mining đó chính là y học và sức khỏe cộng đồng.
* Bài báo ngày cho thấy một cuộc khảo sát các kĩ thuật hiện tại của KDD, sử dụng các công cụ đã tích hợp Data mining trong y tế và sức khỏa công đồng. Ngoài ra còn bao gồm các vấn đề quan trọng và thử thách trong việc kết hợp Data mining với y tế nói chung.
* Bài nghiên cứu còn cho thấy sự phát triển của một lượng lớn các ứng dụng có Data mining, bao gồm các ứng dụng như phân tích dữ liệu tại các bệnh viện để đưa ra các chính sách hoạch định sức khỏe tốt hơn, phát hiện dịch bệnh, ngăn ngừa các trường hợp tử vong trong bệnh viện và phát hiện các gian lận trong quyền bản hiểm.

1. Giới thiệu và các nhân tố cơ bản:

* Mục tiêu:
* Liệt kê các lợi ích và nổi bật tầm quan trọng của Data mining trong y học vào sức khỏe cộng đồng.
* Tìm kiếm các kỹ thuật Data mining đã được sử dụng trong các lĩnh vực khác có thể áp dụng cho sức khỏe cộng đồng.
* Xác định các vấn đề và thử thách trong ứng dụng Data mining vào y khoa.
* Vạch ra những khuyến nghị cho việc thu thập các tri thức trong các cơ sở dữ liệu điện tử thông qua Data mining

1. Phương pháp luận:

* Dựa theo các phương pháp bị hạn chế, phương pháp luận chính được sử dụng trong bài nguyên cứu này thông qua các bài đã được xuất bản trong nhiều lĩnh vực của y học, khoa học và kỹ thuật máy tính. Bài nghiên cứu dựa trên những xuất bản gần đây nhất, từ năm 2000 trở về đây.

1. Các nghiên cứu:
   1. Data mining trong phát hiện dịch bệnh:

* Sử dụng dữ liệu thiết thực để hỗ trợ cho việc ra quyết định trong y học ( còn được biết là EBM – Evidence Based Medicine) đã tồn tại trong nhiều thế kỷ. John Snow, đã được biết đến là cha đẻ của dịch tễ học hiện đại, đã sử dụng các bản đồ đồ thị vào năm 1854 để phát hiện ra nguồn bệnh thổ tả và đã chứng minh rằng bệnh này lây lan qua các hệ thống cấp nước.
* Snow đếm số lượng các bệnh nhân và vẽ sơ đồ địa chỉ của các bệnh nhân trên bản đồng là các thanh màu đen. Ông đã phát hiện rằng hầu hết các ca tử vong đều nằm xung quanh một điểm bơm nước nhất định tại London.
* Florence Nightngale đã phát minh ra biểu đồ vùng phân cực vào năm 1855 để chi ra nhiều ca tử vong của binh lính là do sự mất vệ sinh trong khám lâm sàn và có thể ngăn ngừa được. Bà đã sử dụng những biểu đồ để thuyết phục các chinh sách thực hiện cải cách và đã giảm được số lượng các ca tử vong.
* Snow và Nightingale đã chính mình thực hiện thu thập dữ liệu, sang lọc và phân tích thông qua các dữ liệu về tỉ lệ tử vong trong suốt thời gian nghiên cứu vì số lượng bộ thông tin có thể quản lý được. Ngày nay, số lượng dân số, dữ liệu điện tử thu thập được cùng với sự toàn cầu hóa và tốc độ phát bệnh của các dịch bệnh làm cho việc thao tác dữ liệu gần như hoàn toàn không thể hoàn thành.
* Đó là nguyên nhân tại sao Data mining trở nên hữu ích trong việc chăm sóc sức khỏe. Tuy phát triển khác chậm nhưng vẫn được áp dụng để giải quyết các vấn đề khác nhau của việc thu thập tri thức trong lĩnh vực này.
* Data mining và các ứng dụng của nó trong y học và sức khỏe cộng đồng là một lĩnh vực học tập còn non trẻ. Vào năm 2003, Wilson et al bắt đầu thăm dò các trường hợp mà KDD và kỹ thuật Data mining đã được áp dụng trong cơ sở dữ liệu y tế. Người ta cũng đã phát hiện nhiều sự nhầm lẫn trong các lĩnh vực được tạo thành từ Data mining. “Một vài tác giả cho rằng Data mining chỉ là một quá trình thu thập, trong khi các tác giả khác cho rằng Data mining sử dụng kỹ thuật thống kế trong quá trình thu thập trị thức.” (Wilson et al 2003)
* Vì những quan niệm sai lầm vẫn tiếp tục trong cộng đồng y khoa về những gì mà Data mining chứa đựng, chúng ta sẽ định nghĩa lại những quan niệm đó. Một định nghĩa được chấp nhận nhiều nhất về Data mining ngày nay là một bộ các quy trình và kỹ thuật trong khai phá và mô tả kiểu và hướng trong dữ liệu (Witten and Frank 2005). Chúng ta sẽ sử dụng định nghĩa này xuyên suốt trong bài cáo
  1. Tầm quan trọng và lợi ích của Data mining trong y học và sức khỏe cộng đồng:
* Mặc dù sự khác biệt và xung đột trong hướng tiếp cận, lĩnh vực y tế hiện nay cần nhiều Data mining hơn. Có 1 vài đối số có thể được nâng cao để hỗ trợ trong việc sử dụng Data mining trong lĩnh việc y tế. Không chỉ trong y tế cộng đồ mà còn trong lĩnh vực y tế tư nhân.
* Tràn dữ liệu (*Data overload*). Nhiều tri thức có giá trị được thu bởi việc tin họa hóa các dữ liệu y tế. Tuy nhiên, một lượng lớn dữ liệu được lưu trữ sẽ làm cho việc này trở nên khó khăn và gần như là không thể nếu người ta phải sàng lọc và khai phá tri thức (Cheng, et al 2006).
* Trong thực tế, một số chuyên gia tin rằng những đột phá trong lĩnh vực y tế đã chậm lại, thêm vào đó là độ phức tạp của các thông tin y tế hiện nay. Máy tính và Data mining là thích hợp nhất cho mục đích trên. (Shillabeer and Roddick 2007).
* Y học dựa trên thông tin và ngăn chặn các sai sót trong bệnh viên. Khi các tổ chức y tế ứng dụng Data mining vào dữ liệu, họ có thể khai phá được các tri thức hữu ích và khả năng cứu sống nhiều mạng sống vẫn còn nằm trong cơ sở dữ liệu của họ. Ví dụ, một nghiên cứu đang được thực hiện trong các bệnh viên đã cho thấy rằng 87% ca tử vong trong các bệnh viên tại Mỹ có thể ngăn ngừa nếu các nhân viên bệnh viện (kể cả bác sĩ) cẩn thận hơn trong ngăn ngừa sai sót ( HeathGrades Hospitals Study 2007). Bằng việc khai phá các dữ liệu bệnh viên, các vấn đề về an toàn có thể được phát hiện và giải quyết bởi các quản lý bệnh viện và chính phủ.
* Hoạch định chính sách trong y tế cộng đồng (*Policy-making in public heath*). Lavrac et al. (2007)đã kết hợp GIS và Data mining với nhau, Weka với J48, để phân tích điểm giống nhau giữa các trung tâm y tế tại Slovenia. Sự dụng Data mining, họ có thể khai phá được các mô hình trong trung tâm y tế để có thể đưa ra các chính sách khuyến nghị đến Viện Y Tế Cộng Đồng. Họ đã kết luận rằng “Data mining và hệ thống hỗ trợ ra quyết định, kể cả các phương pháp trực quan mới, có thể nâng cao hệ thống ra quyết định.
* Các yếu tố trên đã nhắc chúng ta nhớ tới một sự cố tại trung tâm y tế Rizal của Pasig City của Indonexia vào tháng 10 năm 2006. Thất bại trong việc thực hiện vệ sinh nghiêm ngặt và các biện pháp khử trùng trong bệnh viên đã góp phần dẫn đến các ca tử vong của trẻ so sinh do bị nhiểm khuẩn. Không ai có thể biết được chính xác chuyện gì đang xảy ra cho tới khi các ca tư vong trở nên thường xuyên. Sau khi kiểm tra dữ liệu của bệnh viên, bộ phận y tế ( Department of Health – DOH) đã tìm thấy 12 trong 28 trẻ được sinh ra vào ngày 4 tháng 10 đã tử vong vì nhiểm khuẩn (Tandoc 2006). Với cơ sở dữ liệu đã được xử lý và một ứng dụng Data mining, DOH có thể các hiện các sự kiện không bình thường và hạn chế chúng diễn biến tệ hơn.
* Tiết kiệm tiền và chi phí. Các tổ chức dùng Data mining để rút trích nhiều cơ sở dữ liệu với chi phí thấp nhất. KDD và Data mining đã được áp dụng vào việc tìm ra các gian lận trong thẻ tín dụng và bảo hiểm (Kou et al. 2004). Thêm vào đó, những kỹ thuật này có thể được sử dụng để tìm ra các bệnh nhân khác biệt trong bảo hiểm y tế - một hệ thống bảo hiểm y tế thuộc quốc gia Philippin đã từng thực hiện điều này.
* Phát hiện sớm và ngăn chặn bệnh dịch. Cheng, el al đã đưa ra dẫn chứng trong việc sử dụng các thuật toán phân lớp để phát hiện các trường hợp mắc bệnh tim, một loại bệnh được quan tâm nhất trên thế giới. Cao et al (2008) đã mô tả việc sử dụng Datamining như một công cụ hỗ việc kiểm soát các thí nghiệm vac xin lâm sàn. Sử dụng Data mining đã giúp cho y học có thề phát hiện ra nhiều bệnh nhân có biểu hiện khác thường hơn là chỉ nhìn vào một tập dữ liệu.
* Ngăn ngừa, quản lý bệnh dịch và đưa ra các chính sách y tế cộng đồng. Các chuyên gia y tế đã bắt đầu tìm cách ứng dụng Data mining vào việc phát hiện và quản lý dịch bệnh. Kellogg et at (2006) đã nêu ra một kỹ thuật kết hợp mô hình không gian, mô phỏng và khai phá dữ liệu không gian để tìm ra các được điểm của dịch bệnh. Phân tích kết quả của việc khai phá dữ liệu để đưa ra các chính sách nhằm phát hiện và quản lí dịch bệnh bùng phát.
* Wong et al. (2005) đã giới thiệu WSARE, một thuật toán phát hiện dịch bệnh khi vừa ở giai đoạn đầu. WSARE – “What’s Strange About Recent Events” được tạo nên dựa trên các luật kết hợp và mạng Bayesian. Áp dụng WSARE trên các mô hình giả lập có thể cho ra những kết quả tương đối chính xác các dịch bệnh đã được mô phỏng. Và nó cũng đi kèm với một lời cảnh báo khi áp dụng thật toán này trong đời sống thật.
* Chuẩn đoán và các hệ thống ra quyết định không gây hại. Một số chuẩn đoán và thí nghiệm có thể gây hại cho bệnh nhân. Lấy một ví dụ là việc tiến hành sinh thiết ở phụ nữ nhằm phát hiện bệnh ung thư cổ tử cung. Thangavel et al (2006) đã sử dụng thuật toán gom cụm K-means để phân tích các bệnh nhân ung thư cổ tử cung và đã nhận thấy việc gom cụm dữ liệu có để đưa ra những kết quản dự đoán tốt hơn các phương pháp hiện có. Ông cũng đã tìm thấy một số bộ thuộc tính có thể được bác sĩ sử dụng như một sự hỗ trợ trong việc ra quyết định đưa các bệnh nhân có khả năng bị bệnh ung thư đi tiến hành sinh thiết hay không.
* Gorunescu (2009) đã mô tả cách sử dụng hệ hỗ trợ chuẩn đoán bằng máy tính (CAD) và siêu âm nọi soi Elastography (EUSE) được ứng dụng Data mining để đưa ra những quyết định ngăn ngừa bệnh ung thư mà không gây hại cho bệnh nhân. Trong khi đó cách truyền thống là bác sĩ sẽ nhìn vào các bộ phim siêu âm và thiết bị để đề nghị bệnh nhân là sinh thiết hay không.
* Quyết định của bác sĩ là hoàn toàn chủ quan, phụ thuộc hoàn toàn vào các Video siêu âm. Gorunescu đã giải quyết việc này bằng một cách hoàn toàn khác, sử dụng Data mining. Ông đã không học về nhân khẩu học của bệnh nhân, thay vì tập trung vào các Video siêu âm, trước tiên họ huấn luyện một thuật toán phân lớp sử dụng MLP (multi-layer percetron) trên các trường hợp khối u lành tính và ác tính.
* Mô hình của họ phân tích các điểm ảnh và RGB của chúng để tìm ra các bệnh nhân nằm phân biệt khối u lành tính và ác tính. Sau đó, họ áp dụng kết quả của mô hình vào các trường hợp khác. Và kết quả của mô hình đã đạt độ chính xác cao trong chuẩn đoán với độ lịch chuẩn rất thấp.
* Phân loại các thuốc có hại (ADEs). Một số thuốc và hóa chất đã được đánh giá rằng sẽ không gây hại cho con người nhưng sau một thời gian dài sự dụng thì kết quả là ngược lại. Wilson et al (2003) đã tiết lộ rằng tổ chức US Food và Drug Administration đã sử dụng data mining để tìm ra các loại thức có hại khi sử dụng lâu dài trong cơ sở dữ liệu của họ. Thuật toán này có tên là MGPS – Multi-item Gamma Poisson Shrinker đã tìm ra 67% loại thuốc có hại sớm hơn 5 năm nếu dùng cách thông thường.
* Chúng ta đã thấy được những ứng dụng của data mining trong việc phát hiện bệnh dịch, ngăn ngừa tử vọng, cải thiện chuẩn đoán và thậm chí phát hiện các gian lận trong bảo hiểm y tế. Dù sao, có nhiều cách sử dụng Data mining khác nhau trong lĩnh vực y tế.

1. Các vấn đề và thử thách:

* Việc áp dụng Data mining vào lĩnh vực y tế gặp rất nhiều khó khăn do chính đặc trưng riêng của lĩnh vực này. Shillabeer và Roddick (2007) đã nêu ra một số mâu thuẫn giữa các phương pháp khai phá dữ liệu truyền thống và y học.
* Trong nghiên cứu, khai phá dữ liệu bắt đầu với một giả thuyết và sau đó đưa ra một kết quả đã được điều chỉnh phù hợp với giả thuyết đó. Trong khi đó, phương pháp chuẩn của khai phá dữ liệu trước đây là bắt đầu với các bộ dữ liệu.
* Ngoài ra, trong khi khai phá dữ liệu truyền thống quan tâm đến mô hình và hướng của các bộ dữ liệu thì trong y học lại ít quan tâm hơn. Khác biệt lớn nhất đó là phương pháp tiếp cận của khai phá dữ liệu chuẩn chỉ quan tâm đến việc mô tả chứ không phải giải thích mô hình và hướng của dữ liệu. Y học thì lại quan tâm đến từ những giải thích vì một khác biệt nhỏ nhặt nhất có thể là ranh giới giữa sống và chết.
* Lấy một ví dụ, bệnh than và bệnh cúm có cùng những triệu chứng về đường hô hấp. Nếu loại bỏ đi một số đặc điểm nhỏ trong khi thực hiện khai phá dữ liệu thì người ta có thể cho đó chỉ một đợt dịch cúm nhưng thật sự đó lại là dịch bệnh than (Wong et al 2005). Không phải ngẫu nhiên mà các bài báo về khai phá dữ liệu trong bệnh dịch và chữa trị luôn đưa ra những kết luận mơ hồ. Những thất bài này cho ta thấy được độ thiếu tin cậy của khai phá dữ liệu trong y tế.
* Việc nhầm lẫn về định nghĩa của khai phá dữ liệu cũng làm vấn đề trở nên phức tạp. Ví dụ, ta có thể thấy một số bài báo với từ khóa là “data mining” nhưng thực sử chúng chỉ là những biểu đồ bình thường. Shillabeer (2009) cho biết việc hiểu làm này là rất thường xuyên đối với những nơi mới bắt đầu áp dụng data mining vào y học.
* Cho dù các kết quả thu được từ khai phá dữ liệu là đáng tin cậy nhưng việc thay đổi thói quen của các bác sỹ, y tá không phải là một điều dễ dàng. Ayres (2008) đã cho biết đã có một số trường hợp các bác sỹ từ chối thay đổi các chính sách của bệnh viện. Đã có một trường hợp một bác sỹ đã quên rửa tay sau khi khám nghiệm tử thi và gây ra một nhiều ca tư vong cho nhiều bệnh nhân được khám ngay sau đó. Đến lúc này, các bác sỹ mới thay đổi các thói quen của họ.
* Shillabeer (2009) cũng cho biết rằng hầu hết các bác sỹ (ở Úc) thường nghe lời khuyên từ các chuyên gia y tế hơn là xem kết quả thu được từ khai phá dữ liệu.
* Ngoài ra những dữ liệu riêng tư của bệnh nhân cũng là một vật cản lớn trong việc ứng dụng khai phá dữ liệu vào y học vì để đưa ra một kết quả chính xác nhất thì cần một lượng lớn những dữ liệu cần thiết. Nhưng chỉ có những dữ liệu này mới có thể giúp con người ta tránh được những bệnh chết người.

1. Kết luận và khuyến nghị:

* Bài nghiên cứu về ứng dụng của khai phá dữ liệu trong y học và y tế cộng đồng đã cho ta thấy sơ qua những thành công và thử thách của khai phá dữ liệu. Các tổ chức sức khỏe có thể ứng dụng chúng để tìm kiếm những tri thức cần thiết nằm trong dữ liệu của họ.
* DOH có thể tìm ra được các trường hợp tử vọng có thể xảy ra để có thể đưa ra các chính sách tốt hơn về kế hoạch tim phòng, xác định vật truyền bệnh sốt rét, ngăn chặn các lỗi trong bệnh viện, dự đoán nơi dịch bệnh sẽ bùng phát và ngăn chặn các gian lận trong bảo hiểm.
* Trước khi bắt tay vào áp dụng data mining, các tổ chức phải thực thi các chính sách về quyền riêng tư vả bảo mật dữ liệu của các bệnh nhân đến các đối tượng liên quan, các chi nhánh và cơ quan của tổ chức.
* Chất lượng dịch vụ của y tế công đồng hoàn toàn phụ thuộc vào việc thao tác và ứng dụng khai phá dữ liệu lên các bộ dữ liệu.

1. **NEXT**