Health Care Data and Analytics 3570

Fall 2023, Wednesdays 10:15-11:45, 0.5 CU
Classroom CPC Auditorium

Instructor: Professor Marissa King (mdking@wharton.upenn.edu)
Teaching Assistants: Ali Bray (alibray@wharton.upenn.edu)
Jackson Reimer (reimerj@wharton.upenn.edu)

Course Overview: Health care data creates unparalleled opportunities to save lives, improve health, strengthen the health care workforce, reduce costs, and increase efficiency. But it also presents a unique set of challenges ranging from privacy to data consistency. In this course, we begin by surveying the health care data landscape and then turn to how to use this rich data to better manage care and organizations. As we work through key challenges, we will refine the art of asking good questions and gain first-hand experience applying analytics to answer them. We will also examine innovative businesses focused on health care data and analytics.

At the end of this course, students will: (1) Understand the topography of the health care data landscape, (2) Have the skills necessary to be thoughtful consumers of evidence on health care, health reporting and health policy debates, (3) Be able to use and interpret the use of analytics intended to improve care and health care management, (4) Anticipate business opportunities in health care data and analytics, and (5) Know where the greatest opportunities to use AI to improve health care exist.

In this class, you will learn how to use data and analytics to improve the business of health care. Identifying important health care problems and translating answers into solutions requires domain expertise, knowledge of available data (as well as its limitations), and analytic vision. This course aims to hone and develop your skills in each of these areas. This course combines brief lectures with live cases based on existing health care problems and gives you the opportunity to solve them by working on real-world health care datasets. We will also learn from leading practitioners how to translate solutions into business opportunities.

Prerequisites: This course assumes familiarity with regression. You should know how ordinary least squares regression works in so far as being able to read and interpret regression coefficients and p-values. The course will have problem sets and in class analyses. These exercises will require that you understand and produce basic summary statistics and regression output.

Coding experience is helpful but not essential. A willingness to read and work with code is required.

This is an *applied* data and analytics course. The focus of this course is on applying analytic tools to solve health care problems, rather than providing an in-depth understanding of how the statistical tools work. Teaching programming is also outside the scope of this course. Students with advance programming or data science experience are welcome but it is essential to realize that the goal of this course is application.

Course Overview:

Most sessions are oriented around a broad health care challenge. We will then use real-world health data and analytic tools to address one example problem within this domain. After understanding how one could tackle an example problem with data and analytics, we will look at the state of the art by examining a business case that uses a similar approach.

Date	Health Care Topic	Data	Analytic Tool	Live Case
8/30	Good questions	Overview	Decoding HC data	Your questions
9/6	Risk stratification	Claims	Electronic phenotyping	Opioid prescribing and NarxCare
9/13	Quality of care	Claims	Regression	Medication for OUD and private equity
9/20	Improving operations	EHR	Logistic regression	ED wait times and readmission
9/27	AI and health care	EHR	Machine learning	Revisit previous case with ML
10/4	Evidence into action	Claims	ML and causality	Independence Blue Cross-Guest speaker
10/11	Wearables	Sensor	Continuous monitoring	KardiaMobile
10/18	Access to care	Claims	RCTs and RDDs	Oregon Health Insurance Experiment
10/25	Diagnostic support	Images	AI business models	Aidoc-Guest speaker
11/1	Patient experience	Text	NLP	Chat GPT
11/8	Communication and coordination	EHR Claims	Network Analysis	Physician collaboration networks
11/15	Applied analytics	EHR	Your choice	NeuroBlu—Guest speaker
11/29	Ethics, Bias, and Privacy			
12/6	Presentations			

Grading, Assignments, and Due Dates

Problem Sets--.You will need to complete three problem sets for this course. You may get help from classmates or teaching assistants with coding and analyses but the work you submit must be your own and name any such collaborations, noting the nature of the collaboration and who did what. All reflection and application components must be completed individually.

Problem sets will be provided with code in R. To complete the problem sets, you will not need to be able to write code but will need to be able to run and make adjustments to the provided code. For individual problem sets, this means that someone can help you figure out to make these adjustments, but you need to run and implement changes to the code yourself. There will be a lot of support and optional tutorials to help you complete the problems sets in R.

You are welcome to use any programming language you are comfortable with to complete the problem sets. The objective is for you to arrive at the correct solution or meaningful insight, not to teach you how to code. If you choose to use a language other than R, you may well be on your own to complete the problem sets.

If you plan to use R but are unfamiliar with it, it is highly recommended that you familiarize yourself with the basics prior to the course. A good resource for doing so is Wharton's AI and Analytics for Business online Intro to R Bootcamp module https://aiab.wharton.upenn.edu/students/online-modules/. There are innumerable free resources on R available online and in print.

The popularity of languages changes frequently, different organizations have different preferred languages, and different languages can be more or less useful depending on the problem at hand. In

class exercises will be demonstrated using a variety of languages. For in class exercises, you will not be responsible for coding, making edits to, or running the code. Instead, you will need to know how to formulate models and understand output.

Final Project--.The final project for this course will allow you to apply the tools and insights from class to an area of health care that is of personal interest. This will also be an opportunity for your classmates to learn about the broader health care data and analytics landscape.

You have three options for the final project: (1) you can use data and analytics to gain insight into a pressing health care problem, (2) propose a new startup or initiative, or (3) evaluate an existing company and devise a strategy for implementation, commercialization, and/or growth.

Projects that use data and analytics to gain insight into a pressing health care problem should include:

1. Clear articulation of the research question
2. Case for why the question is important
3. Description of the data and methods used
4. Results
5. Limitations of the analysis
6. Discussion of how the results can inform business and/or policy

Projects that propose a new startup or initiative or to evaluate an existing company and devise a strategy for implementation, commercialization, and/or growth should include:

1. Clear articulation of the problem—What is the business problem you/they are trying to solve? Why is it important? What is the use case?
2. Explanation of how data and analytics can help solve the problem
3. Describe the types of data and tools necessary to address the problem effectively
4. Where do the greatest opportunities for implementation, commercialization, or growth lie?
5. Discuss forseeable obstacles and limitations and propose realistic strategies to overcome them based on lessons from the course or outside sources
6. Strategic plan

If you chose to propose a startup or devise a strategy for an existing company*, you will need to speak with key stakeholders*. You should plan on conducting a *minimum of six stakeholder interviews*.

A principal (but often unspoken) evaluation criteria for open-ended final assignments such as this is that they should entertain your audience. A fruitful way to accomplish this is by thinking critically about what *you* find exciting about your project and then making sure that this is evident throughout your project's development. Please see grading rubric for dimensions on which your final project will be evaluated.

The group final project will be 30% of your grade. You will be graded on the presentation (5%) as well as your final report (25%). The final report should be no more than 5 pages of single-spaced text. It may include up to 3 pages of figures, tables, and exhibits in addition to the text. For

presentations, we will use a format similar to poster sessions at conferences rather than a more formal power point presentation.

Attendance and Participation--. Attendance and participation are essential for this course. Care has been taken to give a reasonable amount of homework to ensure that you are able to complete all of the assignments. This course is front loaded so that pre-reading and problem sets are due earlier in the course. This is to ensure that your team has time and necessary expertise to work on your final project.

It is expected that you will attend all sessions of this short course. Multiple unexcused absences will result in a failing grade. If you anticipate needing to miss multiple classes, you should not enroll in this course. If an emergency arises, please email the TA and the professor to discuss the possibility of an excused absence and makeup work, which will be given on a very limited basis.

Beyond simply being present, you should come to each class prepared. This class has a no device policy. Please do not use your laptop or phone in class unless instructed by the professor. An extensive body of research has demonstrated that even the presence of a device impedes learning and connection. If there are extenuating circumstances that necessitate your use of a device in class, please reach out to the professor and TA prior to the first class.

Classes will often include mini-cases that will require your in class group to submit responses or analyses plans to problems posed in class. These exercises will count towards your participation grade.

We are very fortunate to have many experts who are graciously sharing their time and expertise with us. It is essential that you give them the respect that they deserve.

Your participation grade will reflect attendance, session engagement, in class exercises, respect for guests and fellow class members, and the extent to which you follow the norms established for this course.

Summary of Grading, Assignments and Due Dates

Assignment Due	**Points**	**Due Date**
Class Survey		9/6
Problem Set 1	15	9/13
Problem Set 2	15	9/27
Team Project Proposal		10/4
Problem Set 3	15	11/8
Team Final Project Write Up	30	12/8
Participation and In-Class Exercises	25	
Total Possible Points	100	

Please note due dates for problem sets to ensure you have sufficient time to compete them. Problem sets will be posted at least one week before the due date. See below for weekly pre-readings. All assignments above are due by 9am on the due date via Canvas.

Office Hours

4:30-5:30 on Mondays (Colonia Penn Center Chestnut Room)
1:30-2:30 on Tuesdays (Colonia Penn Center Faculty Lounge)

On Monday and Tuesday before problem sets are due, office hours will be run as a recitation and will be used to walk through the problem set. Please arrive on time for problem set sessions.

Moderate use of Generative AI permitted:

You may use generative AI programs (e.g., tools like ChatGPT) to for help with coding portions of problems sets and coding portions of the final project. Please note that data use agreements may prohibit the uploading of data to generative AI models. For instance, the MIMIC data we will use in problems sets 2 and 3 prohibits uploading of data to generative AI programs. It is your responsibility to ensure compliance with all data use agreements. Use of generative AI programs for data interpretation and/or reflection questions is prohibited (for your own good). You should note that the material generated by these programs may be inaccurate, incomplete, or otherwise problematic. Beware that use may also stifle your own independent thinking and creativity. You may not submit any work generated by an AI program as your own. If you include material generated by an AI program, it should be cited like any other reference material (with due consideration for the quality of the reference, which may be poor). Any plagiarism or other form of cheating will be dealt with severely under relevant Penn policies. Please do not hesitate to reach out if you have any questions about this policy.

Session 1: Introduction to the Healthcare Data Universe (August 30th)

Live case: Your questions

Topics:

- Introduce data and analytic process
- What is a good question
- Overview of HC data
- File claims and electronic health records
- Decoding HC data
- Promise and perils of HC data

NEJM Catalyst. 2018. "Healthcare Big Data and the Promise of Value-Based Care." *New England Journal of Medicine Catalyst* https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0290

Nguyen, Andrew. "Introduction to Healthcare Data" (Chapter 1). Hands on Healthcare https://www.oreilly.com/library/view/hands-on-healthcare-data/9781098112912/ch01.html

Simonite, Tim. 2022. "When it Comes to Health Care, AI has a Long Way to Go. *Wired* https://www.wired.com/story/health-care-ai-long-way-to-go/

Session 2: Patient Segmentation and Risk Stratification (September 6th)

Live case: Prescription opioids and high-risk prescribing

Topics:

- Patient segmentation
- Risk stratification
- Strengths and weaknesses of file claims
- Electronic phenotyping
- Descriptive statistics
- Policy and business applications: Prescribing limits and NarxCare

Vuik et al. 2016. "Patient Segmentation Analysis Offers Significant Benefits For Integrated Care and Support." *Health Affairs* 35:3. https://doi.org/10.1377/hlthaff.2015.131

Szalavitz. 2021. "The Pain Was Unbearable. So Why Did Doctors Turn Her Away? *Wired* https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/

Data exploration: Understanding the structure of data and what is possible is critical for analysis. To familiarize yourself with claims data, we have posted a sample of the Centers for Medicare and Medicaid Services (CMS) *Linkable 2008–2010 Medicare Data Entrepreneurs' Synthetic Public Use File.* Please spend 30 to 45 minutes looking at the data overview and dictionary and taking a look at the data available.

Session 3: Quality of Care (September 13th)

Live case: Buprenorphine prescribing and private equity

Topics:

- Measuring quality
- Time
- Regression analysis
- Categorical variables
- Policy and business application: Private equity

Weiland, Noah. 2023. "More Doctors Can Now Prescribe a Key Opioid Treatment. Will It Help?" *New York Times*. https://www.nytimes.com/2023/03/03/us/politics/buprenorphine-opioid-addiction-treatment.html

Abelson, Reed and Margot Sanger-Katz. 2023. "Who Employs Your Doctor? Increasingly, a Private Equity Firm." *New York Times* https://www.nytimes.com/2023/07/10/upshot/private-equity-doctors-offices.html

Assignment Due: Problem Set 1 due by 9 am on Sept 13th via Canvas.

Session 4: Improving Operations (September 20th)

Live Case: Length of stay and 72 hour returns in the Emergency Department

Topics:

- Logistic and other types of regression
- Model selection
- Risk adjustment
- Common biases in health care data
- Policy and business application: ED wait times and super utilizers

Janke et al. 2023. "Patients are Dying in Emergency Department Waiting Rooms." https://www.medpagetoday.com/opinion/second-opinions/103166

Weintraub and Zimmerman. 2017. "Fixing the 5 Percent." *The Atlantic.* https://www.theatlantic.com/health/archive/2017/06/fixing-the-5-percent/532077/

James et al. 2021. "Logistic Regression." *An Introduction to Statistical Learning*. New York: Springer., pp. 129-139

If you are rusty on OLS regression, please read/review one of the following:

James et al. 2021. "Chapter 3: Linear Regression. *An Introduction to Statistical Learning*. New York: Springer.

For an intuitive overview: Frost. 2019. *Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models*. pp. 16-51 and 54-63

Session 5: Introduction to AI and Machine Learning in Healthcare (September 27th)

Topics:

- Structured versus unstructured data
- Bias-variance tradeoff
- Regularization
- Overview of machine learning tools
- Barriers and possibilities for ML in health care

Assignment Due: Problem set 2 due by 9am on September 27th via Canvas.

James et al. 2021. "Chapter 2: Statistical Learning." *An Introduction to Statistical Learning*. New York: Springer.

Session 6: Evidence into Action (October 4th)

Live Case: Independence Blue Cross

Guest speaker: Aaron Smith-McCallen, Director, Data Science & Health Care Analytics, Independence Blue Cross

Topics:

- Causal inference
- Cost
- Utilization
- Using ML to inform business and policy interventions

Assignment Due: Your group should submit a short proposal (no more than 250 words) for your final project. Please submit one per group.

Rajpurkar et al. 2023. AI in Health and Medicine. *Nature Medicine* https://www.nature.com/articles/s41591-021-01614-0

Ukert, B, David, G, Smith-McLallen, A, Chawla, R. Do payor-based outreach programs reduce medical cost and utilization? *Health Economics*. 2020; 29: 671– 682. https://doi.org/10.1002/hec.4010

Session 7: Wearable sensors (October 11th)

Live Case: Kardia Mobile

Guest Speaker: David Albert, Founder and Chief Medical Officer at AliveCor

Topics:

- Physiological data
- Establishing an evidence base
- FDA approval process
- Data and clinical integration

Familiarize yourself with Kardia Mobile https://www.kardia.com/

Listen to: Babbage from the Economist: An App a Day Keeps the Doctor Away

https://podcasts.apple.com/sa/podcast/babbage-an-app-a-day-keeps-the-doctor-away/id508376907?i=1000560370337

Session 8: Access to Care (October 18th)

Topics:

- Causality
- Regression discontinuity
- Randomized control trials

Varian, Hal. 2016. "Causal Inference in Economics and Marketing." *PNAS* https://www.pnas.org/doi/10.1073/pnas.1510479113

Baiker et al. 2013. "The Oregon Experiment: Effects of Medicaid on Clinical Outcomes." *New England Journal of Medicine* https://www.nejm.org/doi/full/10.1056/nejmsa1212321

Session 9: Diagnostic Support (October 25th)

Live Case: Aidoc

Guest speaker: Arvind Kadaba, Chief Financial Officer at Aidoc

Topics:

- ML and imagining
- Business models for AI in health care
- Evidence and adoption
- Workflow integration

Listen to: Transforming Healthcare with Artificial Intelligence with Elad Walach, CEO at Aidoc

https://healthpodcastnetwork.com/episodes/outcomes-rocket/transforming-healthcare-with-artificial-intelligence-with-elad-walach-ceo-at-aidoc/

Gormley. 2022. "Aidoc, an AI Healthcare Startup, Nabs $110 Million Expansion Round." https://www.wsj.com/articles/aidoc-an-ai-healthcare-startup-nabs-110-million-expansion-round-11655373604

Research Aidoc at https://www.aidoc.com/. Please either read a research article based on the platform (https://www.aidoc.com/learn/clinical-studies/) or watch a webinar (https://www.aidoc.com/learn/webinars/) based on your interest

Session 10: Patient and Provider Experience (November 1st)

Live Case: Chat GPT

Topics:

- Text analysis
- Large language models
- Patient experience
- Provider burnout
- AI and emotion

For this class, you will need access to several resources which take time to approve. Please **request access and set up an account for the following by 10/27:**

1. **WRDS (Wharton Research Data Services)** https://wrds-www.wharton.upenn.edu/
2. **n2c2-nlp at Harvard** https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/ Under research purpose, "use in Wharton Healthcare Data and Analytics course."
3. **Chat GPT** https://chat.openai.com/auth/login The free version is fine.

Kolata. 2023. "When Doctors Use a Chatbot to Improve their Bedside Manner." *New York Times*

Lee, Goldberg, and Kohane. 2023. *The AI Revolution in Medicine: GPT-4 and Beyond.* New York: Pearson. pp. 1-64, 99-119, 157-164

Session 11: Communication and Coordination Networks (November 8th)

Live Case: Communication and coordination networks

Assignment Due: Problem set 3 due by 9am on November 8th via Canvas.

Session 12: Applied Analytics (November 15th)

Guest speaker: NeuroBlu

Access to the NeuroBlu platform and brief overview slides will be sent to you on Friday. Please spend some time familiarizing yourself with the NeuroBlu platform and its capabilities.

Session 13: Ethics, Bias, Privacy (November 22nd)

Topics:

- Privacy
- Bias
- Federated Learning

Levi, Ryan and Dan Gorenstein. 2023. "AI in Medicine Needs to be Carefully Deployed to Counter Bias and Not Entrench It." *National Public Radio* https://www.npr.org/sections/health-shots/2023/06/06/1180314219/artificial-intelligence-racial-bias-health-care

Obermeyer et al. 2021. *Algorithmic Bias Playbook* Center for Applied AI at Chicago Booth. https://www.chicagobooth.edu/-/media/project/chicago-booth/centers/caai/docs/algorithmic-bias-playbook-june-2021.pdf

Session 14: Group Presentations