Data Article

Data for modelling US projections of product approvals, patients treated, and product revenues for durable cell and gene therapies

Colin M. Young*, Mark Trusheim, Casey Quinn

NEWDIGS Initiative, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, USA

A R T I C L E   I N F O

Article history:
Received 21 September 2021
Revised 14 December 2021
Accepted 25 January 2022
Available online 3 February 2022

Keywords:
Markov chain Monte Carlo
Clinical trials
Time in phase
Success probability
Durable cell and gene therapies
Incidence and prevalence

A B S T R A C T

The recent marketing approval of several durable gene and cell therapies (2017-2020), together with observations that 7,000 monogenic indications and many cancers were potential targets, led to concern about the potential economic impact of such therapies on the US healthcare system. Using a Markov chain Monte Carlo simulation model, driven stochastically by our estimates of the time in phase of clinical trials and each clinical trial phase probability of success, we forecast the pattern of future US regulatory approvals for such therapies currently undergoing clinical trials. Using parameters of those trials, such as inclusion and exclusion criteria, and other epidemiological data we estimate potential treatable patient populations and use these together with pricing estimates to forecast a range for the potential future list price product revenues associated with these therapies.

© 2022 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

DOI of original article: 10.1016/j.drudis.2021.09.001
* Corresponding author
E-mail address: colyoung@mit.edu (C.M. Young).

https://doi.org/10.1016/j.dib.2022.107891
2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
### Specifications Table

| Subject                      | Health Economics |
|------------------------------|------------------|
| Specific subject area        | Markov chain Monte Carlo simulation of the productivity for the US through 2030 of the current durable cell and gene therapy development pipeline |
| Type of data                 | Table            |
|                              | Image            |
|                              | Chart            |
|                              | Graph            |
|                              | Figure           |
| How data were acquired       | Data were extracted manually from the Pharmaprojects®, clinicaltrials.gov and SEER databases, using purpose written SQL code from the AACT database, and from both gray and academic literature. |
| Data format                  | Secondary data that has been filtered and analysed. Excel files with data have been uploaded |
| Parameters for data collection | Forecasting dataset: All products that (i) employ or modify DNA or RNA, (ii) are administered in a single course, and (iii) are expected to deliver at least 18 months average benefit. We broadly apply the US Food and Drug Administration's (FDA) Center for Biologics Evaluation and Research (CBER) definition for gene therapy products (15). Qualifying therapies were those falling into the following modalities: |
|                              | • Gene replacement therapies both in vivo and ex vivo using viral vectors |
|                              | • T-cell receptors (TCRs) and immune cells engineered to incorporate chimeric antigen receptors (CARs) |
|                              | • Gene editing therapies: |
|                              |   o Zinc finger nucleases (ZFNs) |
|                              |   o Transcription activator-like effector nucleases (TALENs) |
|                              |   o CRISPR-Cas9 (clustered regularly interspaced short palindromic repeats) |
|                              | • Long-acting DNA plasmids. |
| Description of data collection | Clinical trials are identified for these products. Only interventional trials in Phase 1, Phase 2 (including Phase 1/2) and Phase 3 (including Phase 2/3) that are, or could be, currently active (i.e. with the following statuses: Recruiting; Active, not recruiting; Enrolling by invitation; and Completed) are included. |
| Data source location         | Pharmaprojects®[Informa 2021 database [3] |
|                              | AACT- the SQL database format of clinicaltrials.gov [4] |
|                              | SEER – the National Cancer Institute’s Surveillance, Epidemiology, and End Results Program database [5] |
|                              | Academic and gray literature |
| Data accessibility           | Young, C., “Durable Cell and Gene Therapy Potential and Financial Impact”, Mendeley Data, v1, (2021). [2] http://dx.doi.org/10.17632/djm65zrhmt.1 |
| Related research article     | Young C, Quinn C, Trusheim M. Durable Cell and Gene Therapy Potential Patient and Financial Impact: US Projections of Product Approvals, Patients Treated, and Therapeutic Revenues. Drug Discovery Today In Press https://doi.org/10.1016/j.drudis.2021.09.001 |
Value of the Data

- These data are key source material for readers understanding our Markov chain Monte Carlo simulation model and methods
- Any researchers seeking to better understand US clinical pipeline projections will benefit from these data either directly, or using them as an example to follow in any similar data analysis
- We recommend any researchers seeking to replicate or expand our analysis within the US use these data as a source for consistency; we recommend any researchers going beyond this (e.g., to other countries or therapy types) use these data as a reference for the type and source that should be used.

1. Data Description

Parameter Estimation Datasets:

Trials_CARB&TCR; Trials_Other Oncology; Trials_Gene Therapies; Trials_Cellular Therapies.
PED1. Trials_CARB&TCR
PED2. Trials_Other Oncology
PED3. Trials_Gene Therapies
PED4. Trials_Cellular Therapies

These four datasets contain one entry for each trial (or FDA interaction) and are used for estimating success rates and for creating empirical probability distributions for the time spent in each development phase. Each entry contains the unique Pharmaprojects® ID of the product, the clinicaltrials.gov (henceforth Ctgov) identifier for the trial (except for when the line represents a Registration or Approval event), the trial phase, the trial status, start date, primary completion date, and completion date. (Also included for the convenience of the reader are the name of the originating entity and the drug name.) Table 1 provides information on the distribution by phase of the data in each of the four datasets. Within each data set the trials data are ordered first by product (represented by a unique Pharmaprojects® ID), then by Indication (or Condition) and, finally, by development phase. (All referenced datasets and supplementary materials may be found in Young [2].)

Forecasting Datasets:
Clinical Trials Data; Preclinical Data; Genes & Antigens; Incidence, Prevalence, Adoption.
FD1. Clinical Trials Data:

This dataset contains one entry for each clinical development program in the forecasting dataset (typically the trial with the highest phase and the most recent start date). Each

| Early Phase 1 | CAR-T&TCR | Other Oncology | Gene Therapy | Cellular Therapy | Totals by Phase |
|---------------|-----------|----------------|--------------|------------------|-----------------|
| Phase 1       | 310       | 532            | 77           | 283              | 1202            |
| Phase 1 Phase 2 | 167      | 297            | 133          | 253              | 850             |
| Phase 2       | 67        | 325            | 73           | 349              | 814             |
| Phase 2 Phase 3 | 0        | 15             | 6            | 11               | 32              |
| Phase 3       | 10        | 61             | 39           | 80               | 190             |
| Registration  | 8         | 4              | 4            | 22               | 38              |
| Approval      | 5         | 3              | 3            | 17               | 28              |
| Totals        | 626       | 1243           | 335          | 1024             | 3228            |
entry contains the trial’s CTgov identifier, the indication addressed, the antigen or gene addressed, the originator country, the delivery vector, the trials status, phase, start date and end date. Also, where appropriate, the dates of entering registration (being accepted for priority review) and receiving marketing approval. (Also included for the convenience of the reader are the name of the originating entity and the drug name.) The data are ordered by indication, then by antigen or gene targeted and, finally, by delivery vector.

FD2. Preclinical Data:

This dataset contains one entry for each product in preclinical development that is included in the forecasting dataset. Each entry contains the product’s unique Pharmaprojects® identifier, the indication addressed and, where available, the gene targeted, the date the product was added to the Pharmaprojects® database, the originator country and the treatment modality. (Also included for the convenience of the reader are the name of the originating entity and the drug name.) The data are ordered by indication, then by gene targeted and, finally, by treatment modality.

FD3. Genes & Antigens:

This dataset is ordered by indication and contains a line for each gene or antigen associated with the indication that is being addressed by a product in either the Clinical Trials Data or Preclinical Data datasets. Each entry contains the gene or antigen name and the proportion of the total indication population for whom either that gene might be the proximate target of treatment or, in the case of cancers, the proportion that express the antigen targeted.

FD4. Incidence, Prevalence, Adoption:

This dataset is ordered by indication and contains an entry for each indication with an entry in the Genes & Antigens dataset. Each entry contains the Indication, Therapeutic Category, estimates of total indication Incidence and Prevalence, estimates of the likely clinically eligible Incidence and Prevalence and the percentage of the total indication population thus represented. Also included for each indication are the parameters used in a Bass model [7] used to estimate the adoption rate of the therapy, if approved, and the time taken to clear the prevalent population.

Supplementary data files:

Bibliography; SQL code used for accessing the AACT (Aggregate Analysis of ClinicalTrials.gov) database; Clinical trials – inclusion and exclusion criteria

S1.Bibliography:

Contains references, keyed by Indication and, within Indication, by Gene or Antigen, used in estimating the Indication incidences and prevalences found in the Incidence, Prevalence, Adoption dataset.

S2. SQL code used for accessing the AACT database:

Contains examples of the SQL code used to extract data from AACT.

S3. Clinical trials – inclusion and exclusion criteria

Contains participant inclusion and exclusion criteria for clinical trials in our forecasting sample

S4. Methods for processing data extracted from Pharmaprojects®

Contains an Excel workbook (with VBA modules) that processes data extracted from Pharmaprojects® and provides a list of clinical trials identifiers linked with specific therapies (and their citeline® identifiers).
Table 2
Keyword filters used to select products from the Pharmaprojects® database.

| Activity                          | Therapeutic Class                          |
|----------------------------------|-------------------------------------------|
| Biotechnology Products           | 1. Antisense therapy                       |
|                                  | 2. Bispecific T cell engager               |
|                                  | 3. Cellular therapy, chimeric antigen receptor|
|                                  | 4. Cellular therapy, other                 |
|                                  | 5. Cellular therapy, stem cell             |
|                                  | 6. Cellular therapy, T cell receptor       |
|                                  | 7. Cellular therapy, tumour-infiltrating lymphocyte |
|                                  | 8. Gene therapy                            |
|                                  | 9. Lytic virus                             |
|                                  | 10. Messenger RNA                          |
|                                  | 11. Oligonucleotide, non-antisense, non-RNAi|
|                                  | 12. RNA interference                       |

Table 3
Examples of keywords used in CTgov searches

| CAR-T                          | CAR-NK                                  |
|--------------------------------|-----------------------------------------|
| Chimeric antigen receptor      | TCR                                     |
| T-cell receptor                | AAH                                     |
| Adeno-associated virus         | Lenti-virus                              |
| Adoptive cell transfer         | CD34                                     |
| CD34+                          | CD34+ AND Transduce                     |
| Stem cell                      | Transduction                            |
| CRISPR                         | TALENS                                  |
| ZFN                            | Zinc finger                             |

2. Experimental Design, Materials and Methods

Records for gene therapy products and cellular products were extracted manually from the Pharmaprojects® database using hierarchical keyword-based filters (Table 2) and assigned to one of four categories (CAR-T&TCR; Other Oncology; Gene Therapies; Cellular Therapies).

Each product record was processed using Excel and VBA to provide basic identification data and, where available, a list of clinical trials (CTgov identifiers) and the indications that they address. For each CTgov registered clinical trial, status, phase and starting and completion date data were acquired from AACT (database for the Aggregate Analysis of ClinicalTrials.gov), together with confirmatory information on the indications addressed by the trial. Further trials for these therapies, and clinical stage therapies not previously identified, were found in CTgov using a combination of natural language processing and manual searches and extraction (Examples of keywords used in Ctgov searches may be found in Table 3). Additionally, for each identified therapy, dates for “Registration” (acceptance for priority review of a BLA or NDA by the US FDA) or “Approval” were obtained, where available, from Pharmaprojects® or from the publicly available datashares. These data form the basis of the four Parameter Estimation Datasets listed above (PED1, PED2, PED3, PED4). Data included in these datasets were curated to identify candidates for the Forecasting Dataset (FD1. Clinical Trials Data). Additionally, a number of non-oncological, durable, cell and gene therapy products in later stage preclinical development were identified in the Pharmaprojects® extracted data and used to create FD2 (Preclinical Data).
Table 4
Pricing assumptions for durable cell and gene therapies.

| Disease Types            | Cost ($) |
|--------------------------|----------|
| Oncology                 | 400,000  |
| Ultra Orphan             | 1,500,000|
| Orphan                   | 800,000  |
| Higher prevalence        | 500,000  |
| Orphan Ophthalmology     | 800,000  |
| Unit elasticity (high)    | 100,000  |
| Unit elasticity (low)     | 50,000   |

The Supplementary Bibliography, S1, provides reference resources used to develop estimates of the proportions of the total prevalence of an indication that a particular gene is responsible for, in the case of oncological indications, the proportion that express a particular antigen. Those estimates are found in dataset FD3. In many cases antigen expression rates are not well documented. In those cases, placeholders using our best judgment based on the limited evidence in the context of other antigen expression rates are used.

The Supplementary Bibliography also contains the references used, together with the clinical trials inclusion and exclusion criteria found in Supplementary file S3, to develop estimates of incidence and prevalence, and clinical eligibility, for non-oncological indications. Those estimates are found in dataset FD4.

The SEER database was queried manually for each oncological indication referenced in FD1. For each, sixteen years of both incidence by age and five-year survival data were extracted. At the time of extraction, the most recent data available was for 2015. Simple linear forecasting models were applied to each age group for each indication to provide estimates of contemporary data. Five-year survival data was used to calculate five-year mortality (the converse of survival), which was used as our estimate of clinical eligibility for all last line of therapy (relapsed and refractory) oncological indications. As SEER data is collected from states covering only around 27% of the US population, all incidence and clinical eligibility estimates were scaled to reflect the full population. These data are found in FD4. (Note that, for oncological indications, only incidence data is used in the forecasting process – there is essentially no prevalence as the eligible population is made up of terminally ill patients.)

Adoption parameters found in dataset FD4 drive a Bass diffusion model, which captures how ‘innovators’ and ‘imitators’ combine to determine overall adoption of a new technology [7]. Factors considered in assigning model parameters included disease severity (life expectancy, quality of life, and availability today of alternative therapies), economics (cost of existing standard of care therapies and Medicaid or Medicare reimbursement policies), and friction between various stakeholders. Assumptions around adoption parameters were tested and refined within the MIT NEWDIGS FoCUS team and the wider consortium.

For oncological therapies we note that 20% of patients enrolled in acute lymphoblastic leukemia (ALL) and diffuse large B-cell lymphoma (DLBCL) late-stage trials did not receive an infusion [8]. This, together with evidence of considerable regulatory drag in reimbursement setting by CMS led us to apply a peak adoption level of 75% to all oncology treatments. We have further assumed a time to peak of 2 years although, at the time of writing, the revenue performance of the five approved therapies does not support such a compressed timeline.

Finally, Table 4 contains the pricing assumptions used in the companion paper, Young et al. [1]. The prices reflect a combination of observed market prices (may be found in Young et al [1]), economic analysis and assumption. Ultra Orphan refers to indications with a US prevalence <10,000; Orphan indications have a US prevalence < 200,000. We assume that certain therapies will compete with an established (and successful) standard of care (SoC). We assume that in those cases the market price of the new therapy will be highly dependent on the cost of the current SoC. For these therapies we use an assumption of local unit elasticity.
Ethics Statement

This work complies with Elsevier’s duties of authors as described in its publishing ethics policies at https://www.elsevier.com/about/policies/publishing-ethics. This work has not involved human or animal subjects, nor has it used social media data.

CRediT Author Statement

Colin Young: Conceptualization, Methodology, Data Analysis, Writing – original draft preparation; Mark Trusheim: Conceptualization, Methodology, Writing – review & editing; Casey Quinn: Conceptualization, Methodology, Writing – review & editing.

Declaration of Competing Interest

Colin Young reports other from FoCUS Consortium, during the conduct of the study. Mark Trusheim reports other from FoCUS Consortium, during the conduct of the study; other from Co-Bio Consulting LLC, outside the submitted work. Casey Quinn reports other from FoCUS Consortium, during the conduct of the study.

Acknowledgments

We would like to thank all the research assistants whose long and diligent work made this research possible. Particular thanks are due to Harshil Donga, Drs. Jenniffer Bulovic and Yupu Wu, and to the Jonathans, Thomas and Trusheim.

References

[1] Young C, Quinn C, Trusheim M. Durable Cell and Gene Therapy Potential Patient and Financial Impact: US Projections of Product Approvals, Patients Treated, and Therapeutic Revenues. Drug Discov. Today. In Press. doi:10.1016/j.drudis.2021.09.001
[2] C. Young, Durable Cell and Gene Therapy Potential and Financial Impact, Mendeley Data (2021), doi:10.17632/djm65zrhmt.1.
[3] Informa PLC. Pharmaprojects®. https://citeline.informa.com/drugs/results. Accessed 2018-2021.
[4] US National Library of Medicine, ClinicalTrials.gov. https://clinicaltrials.gov/. Accessed 2018-2021.
[5] AACT clinical trials transformative initiative. https://www.ctti-clinicaltrials.org/aact-database. Accessed Feb-Apr 2021.
[6] National Cancer Institute, The surveillance, epidemiology, and end results (SEER) Program. https://seer.cancer.gov/ (2018-2021)
[7] F.M. Bass, A new product growth for model consumer durables, Manag. Sci. 15 (5) (1969) 215–227, doi:10.1287/mnsc.15.5.215.
[8] Institute for Clinical and Economic ReviewVoretigene Neparvovec: Final Report and Meeting Summary, 2018.